

PRODUCTIVE FRICTION AND LEARNER ENGAGEMENT IN GENAI SUPPORTED EFL VOCABULARY LEARNING

MAVERICK VINALES

Independent Scholar, United States

maverick.vinales.scholar@gmail.com

First Received: July 25, 2026

Final Proof Received: Oct 20, 2026

ABSTRACT

This article explains how generative artificial intelligence (GenAI) prompts can support durable vocabulary learning while preserving learner engagement and autonomy in English as a foreign language (EFL) education. A purposive conceptual synthesis integrates vocabulary research, retrieval practice, task involvement, cognitive load, self regulation, and AI literacy. The analysis distinguishes established findings from proposed relationships and develops the AI Mediated Cognitive Engagement Model (ACEM). ACEM locates potential learning benefits in the sequencing of an initial attempt, contingent assistance, evaluative feedback, contextual reuse, and later retrieval. Productive friction is defined prospectively through learning relevant demands, accessible support, and opportunities for error repair; its effectiveness requires independent testing. Four propositions address retrieval, proficiency dependent support, evaluative reuse, and learner control. The synthesis is selective and reports no intervention data. The proposed mechanisms require comparisons that separate prompt sequence effects from exposure, time, feedback quality, and technological novelty. Teachers can use a graduated prompt sequence, verified lexical resources, and equivalent offline tasks to protect participation, reduce unnecessary interface demands, and assess learning after AI assistance is withdrawn. The framework connects prompt design with observable learner activity and delayed lexical outcomes, while treating emotional engagement, teacher judgement, and equitable access as conditions of implementation.

Keywords: EFL vocabulary; generative artificial intelligence; learner engagement; productive friction; prompt design.

I. INTRODUCTION

Generative artificial intelligence (GenAI) changes the timing and availability of assistance in English as a foreign language (EFL) education. A learner can request a translation, an explanation of a collocation, or a model sentence during the moment of uncertainty. This responsiveness creates opportunities for individualised practice, but access to an answer is different from acquiring knowledge that remains available when the tool is absent. Vocabulary learning exposes this distinction particularly clearly: understanding an explanation while reading it does not establish that the learner can later recall the word, discriminate its meanings, or use it appropriately (Aziz, 2024). The instructional problem is therefore to organise appropriate assistance around the lexical activity that the learner needs to perform.

Research on language learning technologies supplies reasons for both interest and caution. Huang et al. (2022) identify pedagogical affordances and implementation challenges across chatbot studies, while Kohnke et al. (2023a)

describe possible uses of ChatGPT for language practice and support. These contributions establish a range of uses rather than a universal learning advantage. The review by Li et al. (2024) identifies needs for longitudinal evidence and closer examination of feedback quality. Li et al. (2025) subsequently show that empirical GenAI research has expanded, while remaining concentrated in higher education and EFL settings, with writing especially prominent. These distributions make a vocabulary specific account useful, but they do not justify assuming that all learners benefit from more frequent chatbot interaction.

Evidence concerning perceived benefits also needs careful interpretation. In their qualitative study, Karataş et al. (2024) document learners' experiences of using ChatGPT, including perceived support for vocabulary, grammar, and writing. Such accounts illuminate acceptability and possible mechanisms; they do not independently establish delayed vocabulary retention. Similarly, a polished AI assisted sentence is evidence about a jointly produced task outcome. It cannot, without additional assessment, demonstrate what its author has learnt. Vocabulary assessment must identify the component of knowledge being tested and provide evidence supporting the intended interpretation of scores (Schmitt et al., 2020). This requirement becomes especially important when an external system can supply the target performance.

The corresponding design dilemma concerns when to provide help. Immediate explanation can be appropriate when an unfamiliar word blocks comprehension or when the learner lacks a useful starting point. Yet an explanation supplied before any attempt can also remove an opportunity for retrieval, contextual inference, or comparison. Research on spacing and retrieval supports arranging practice so that learners reconstruct knowledge and revisit it over time (Carpenter et al., 2022). However, difficulty is not intrinsically beneficial. A demanding task can exceed available knowledge or consume attention without strengthening the intended representation. Comparisons between desirable difficulties and cognitive load theory underline the need to specify the task and the kind of difficulty involved (Pyke et al., 2025).

This article uses productive friction to describe a design intention: introducing a manageable demand that directs attention to a learning relevant operation before or during assistance. Examples include recalling a previously encountered meaning, justifying a choice between collocations, and revising an original sentence after feedback. The term does not mean slow interfaces, repeated compulsory guesses, or withholding necessary accessibility support. Nor is productivity inferred merely because a task feels effortful. Whether a particular demand produces learning is an empirical question that requires independent outcomes. This distinction prevents productive friction from becoming a circular label applied only after an intervention appears successful.

Prompt design matters because it can influence the allocation of work between learner and system. Here, a prompt sequence means the connected instructions and response opportunities through which a task unfolds, including the conditions for requesting help and receiving feedback. Technical competence

in requesting outputs is insufficient on its own. Knoth et al. (2024) connect prompting with AI literacy, while Ma et al. (2024) place prompt use alongside evaluation and ethical understanding in language teachers' ChatGPT literacy. Research with nonexpert prompt designers also cautions against assuming that users naturally test and refine instructions systematically (Zamfirescu-Pereira et al., 2023). Pedagogical sequencing therefore requires explicit modelling and support.

The article proposes the AI Mediated Cognitive Engagement Model (ACEM) for teacher guided EFL vocabulary activities. Its central claim is conditional: a prompt sequence may improve learning when it elicits relevant learner processing, provides dependable assistance, and preserves participation. Three questions organise the argument. Which learner operations should a prompt sequence protect? How can instructional demands be distinguished from unnecessary burden? What observations would support or challenge the proposed relationships? The contribution is an explicit connection between prompt decisions, observable activity, and independently assessed lexical outcomes, supplemented by boundary conditions and falsifiable propositions. This addresses the intersection of AI mediated English education, learner autonomy, engagement, and equitable participation rather than treating technology use itself as the educational outcome.

II. METHOD

Conceptual design and source selection

The article employs purposive theory synthesis and model development. Following Jaakkola's (2020) account of conceptual research design, the theories are selected for their explanatory roles rather than combined into an undifferentiated framework. Vocabulary research specifies the knowledge to be learnt. Retrieval research addresses the timing and reconstruction of that knowledge. The Involvement Load Hypothesis identifies task demands involving need, search, and evaluation. Cognitive load theory constrains demands relative to prior knowledge. Engagement and self regulation research explain participation, monitoring, and help seeking. AI literacy research addresses the additional task of evaluating machine generated assistance.

The selection combines foundational theoretical contributions with recent reviews and directly relevant studies. Sources were retained when they clarified a construct, examined a proposed learning mechanism, exposed an important qualification, or informed implementation in language education. Publisher and institutional records were consulted to verify bibliographic details and the reported nature of the evidence. This is a selective conceptual synthesis, not a systematic review or meta-analysis. It does not claim exhaustive database coverage, a reproducible census of all publications, pooled effects, or an independently coded evidence corpus. These limits affect the breadth of the argument and should be considered when evaluating its contribution.

Analytical procedure and boundaries

The analysis proceeds through three operations. First, constructs are separated by level: prompt features belong to instructional design, engagement describes learner activity and experience, and retention concerns subsequent performance. Second, candidate relationships are examined for compatible explanations and tensions. For example, retrieval opportunities may support retention, whereas unfamiliar instructions may increase burden; neither observation entails that all added difficulty should be encouraged. Third, these relationships are translated into propositions with specified comparisons, boundary conditions, and potentially disconfirming observations. The classroom sequence illustrates how the proposed relationships could be implemented; it is not an account of an observed lesson.

The framework principally concerns deliberate practice with selected English vocabulary after sufficient initial exposure. It is less directly applicable to unrestricted conversation, incidental exposure during extended reading, specialist terminology requiring expert verification, or learners unable to understand the task instructions. Evidence from general cognitive research is treated as a rationale for testing EFL designs, not as proof of their effectiveness. Likewise, findings concerning older chatbots, particular commercial systems, or adult university learners cannot automatically establish effectiveness across current models, ages, institutions, and languages. The synthesis seeks explanatory coherence and testable distinctions within these boundaries.

III. RESULTS AND DISCUSSION

Vocabulary learning and the limits of immediate performance

The first conceptual result is a distinction between successful assistance and independent lexical learning. Knowing a word involves relationships among its form, meaning, and use, and receptive and productive knowledge place different demands on the learner (Nation, 2013). A learner may recognise allocate in a familiar sentence yet fail to retrieve it when explaining a new budget decision. An activity designed around meaning recognition therefore cannot support an unrestricted claim about productive vocabulary development. Webb's (2005) comparison of receptive and productive vocabulary learning provides a basis for matching instructional activity and assessment rather than treating vocabulary as a single undifferentiated score.

The distinction between learning and performance also concerns time. Immediate fluency can be influenced by cues, recent exposure, or external support, whereas learning requires evidence of a more durable change (Soderstrom & Bjork, 2015). Retrieval practice research establishes that actively reconstructing an answer can support later retention, including in vocabulary learning tasks (Karpicke & Roediger, 2008). Feedback matters because an unsuccessful attempt alone does not guarantee a correct representation. Roediger and Butler (2011) discuss retrieval practice in relation to long term retention,

providing a rationale for assessing learning after the immediate benefits of assistance have dissipated. ACEM accordingly separates supported practice from later assessment without access to the chatbot.

EFL vocabulary findings qualify any simple rule that spaced or effortful activity improves every outcome. Nakata and Elgort (2021) found advantages of spaced contextual encounters on explicit vocabulary measures, without a corresponding advantage on their semantic priming measure. Karatas et al. (2025) report benefits from an approach combining retrieval, feedback, and spacing, but the combined treatment cannot isolate the contribution of each component. Uchiyama et al.'s (2019) meta-analysis likewise identifies variation in the relationship between repetition and incidental vocabulary learning. These findings support repeated, assessed opportunities while cautioning against attributing gains from a complex AI activity solely to its prompt wording. They also make outcome selection a theoretical decision rather than a convenient final step.

Friction as a bounded instructional demand

The Involvement Load Hypothesis offers a vocabulary specific account of task demands through need, search, and evaluation (Laufer & Hulstijn, 2001). In an AI mediated activity, receiving an answer may reduce a search requirement while leaving evaluation possible. Conversely, asking several questions may involve extensive interaction without requiring the learner to evaluate lexical alternatives. Yanagisawa and Webb's (2021) meta-analysis indicates that the components do not contribute equally and that moderators matter. ACEM therefore avoids assigning an involvement score from prompt length or treating any extra conversational turn as deeper processing. The task must specify what the learner needs to decide and what evidence demonstrates that decision.

Cognitive load theory supplies a complementary constraint. The demands of a task depend on what information must be coordinated and what the learner already knows; avoidable presentation demands can compete with learning (Sweller et al., 2019). A request to infer an unknown academic word from an equally unfamiliar paragraph may be difficult for reasons unrelated to the intended lexical distinction. Introducing a brief explanation or a familiar context can make the subsequent comparison more accessible. The design objective is to preserve the relevant operation while reducing competing demands. Consequently, ACEM does not prescribe maximal difficulty, a fixed waiting period, or the same number of attempts for every learner.

The model also avoids treating germane cognitive load as a direct synonym for productive friction. Kalyuga (2011) questions the conceptual independence of germane load within the traditional three component account. Although instruments such as that developed by Leppink et al. (2013) provide approaches to perceived load measurement, their use in a new EFL interaction requires validation of the intended interpretation. A higher rating cannot establish that learners retrieved, evaluated, or retained more vocabulary. ACEM instead separates the designed demand, the learner's reported experience, observed

activity, and later outcomes. Agreement among these sources would strengthen an explanation; disagreement would indicate a need to revise it.

Engagement and learner control

Engagement provides the link between an offered task and the learner's actual participation. Hiver et al. (2024) show substantial variation in how language learning engagement has been defined and measured. Philp and Duchesne (2016) discuss cognitive, behavioural, emotional, and social dimensions in language tasks. These dimensions are related but cannot be substituted for one another. A high number of messages may indicate persistence, confusion, or answer seeking. Enjoyment can support continued participation without demonstrating accurate lexical learning. An explanation of a word choice may suggest evaluative activity but still contain an error. ACEM treats such observations as evidence requiring interpretation, not as interchangeable indicators of a single engagement quantity.

Learner control is especially relevant when assistance is deliberately staged. Self regulation includes setting goals, monitoring progress, and adapting strategies rather than simply working alone (Panadero, 2017). Agentic engagement additionally concerns learners' constructive contributions to the instruction they receive (Reeve & Tseng, 2011). Within ACEM, learners can request a hint, choose a familiar context, challenge an explanation, or ask to see a worked example. These actions can improve the fit between a task and a learner's current needs. An option to obtain help is therefore part of the design rather than a failure of the intervention. Research must examine how the option is used and whether regulation improves beyond the scaffolded activity.

Emotional and social conditions can alter whether the intended processing occurs. A learner who experiences a withheld answer as a public test may stop contributing, while a private attempt followed by respectful feedback may preserve willingness to continue. Pair work can make lexical reasoning visible, but it can also allow one participant to perform all the retrieval. These are proposed implementation contingencies rather than claims that one interaction format is universally superior. Teachers should make the purpose of an attempt explicit, permit proportionate help, and distribute opportunities to respond. The resulting design supports engagement through understandable expectations and recoverable errors, while leaving the value of particular arrangements open to investigation.

Architecture and explanatory contribution of ACEM

ACEM organises vocabulary practice into a recurring sequence: orient to the target, attempt a relevant operation, obtain contingent assistance, evaluate feedback, reuse the item, and revisit it later. Orientation identifies the intended knowledge and checks that the learner has a viable starting point. The attempt creates an opportunity for retrieval or inference. Assistance responds to difficulty through a cue, example, explanation, or direct answer. Evaluation requires a decision about meaning or use. Reuse places the item in a different context, and a later encounter tests its availability after the immediate interaction. These stages

are instructional functions; a lesson may combine them or return to an earlier stage.

The proposed explanatory pathway connects prompt sequencing to realised learner operations and then to independently measured vocabulary outcomes. Prior lexical knowledge, proficiency, feedback accuracy, emotional response, and access conditions can modify the pathway. The sequence does not directly create engagement merely by requiring a response. Learners may copy an answer, ignore feedback, or complete a sentence without processing the target distinction. Implementation fidelity must therefore include what the learner actually did. Table 1 separates the core constructs from possible observations and the inferences those observations cannot establish on their own.

Table 1-Constructs and evidence boundaries in ACEM

Construct	Indicative evidence	Inference boundary
Designed demand	Prompt requires recall or comparison before explanation	Requirement does not establish enactment
Cognitive engagement	Learner justifies a lexical choice or revises an interpretation	Explanation may be copied or inaccurate
Emotional engagement	Reported interest, frustration, or willingness to continue	Enjoyment does not establish retention
Regulation and agency	Learner selects support and explains the choice	Help avoidance is not necessarily autonomy
Independent learning	Delayed recall and appropriate use without AI	One test cannot represent all word knowledge

The contribution lies in specifying this allocation and sequence of work within an adaptive dialogue. Retrieval theory explains why reconstruction can matter; involvement theory identifies lexical decisions; cognitive load theory limits unnecessary demands; engagement and regulation theories explain enactment. ACEM connects these functions without proposing that their constructs are identical. A chatbot's ability to respond contingently creates opportunities to alter assistance within a task, while its ability to produce complete answers creates opportunities to bypass the target operation. This tension makes the timing and ownership of each decision analytically important. The model's added value should ultimately be judged by whether these distinctions improve prediction and instructional decisions.

The framework also permits a substantive null result. If a verified worksheet produces equivalent retrieval, feedback, and delayed performance, GenAI may offer convenience or flexibility without an additional learning advantage. If a

chatbot increases interaction but learners perform worse after support is removed, the intended sequence has not achieved its educational purpose. Such outcomes would challenge claims of a distinctive technological benefit while remaining compatible with established learning principles. ACEM therefore does not equate the presence of an AI system with innovation. It directs attention to which instructional functions technology changes, for whom those changes are useful, and at what practical cost.

A further implication concerns the direction of adaptation. Many assistance systems are evaluated by how quickly they remove uncertainty. ACEM instead proposes that adaptation should respond to the purpose of the current episode. During reading for meaning, a direct explanation may preserve comprehension; during deliberate retrieval practice, the same explanation may prematurely supply the target response. The appropriate assistance therefore depends jointly on the learner's state and the instructional goal. Researchers can test this claim by examining whether learners benefit from different support sequences across these two purposes. This distinction also prevents a vocabulary practice routine from being imposed indiscriminately on every encounter with an unfamiliar word.

An illustrative EFL vocabulary sequence

Consider an intermediate EFL class preparing short proposals for a community project. The target item is *allocate*, including its association with distributing resources for a purpose. The teacher first selects an appropriate meaning, verifies examples against a dependable dictionary or corpus, and checks that the surrounding language is familiar. Learners encounter a short statement about a committee distributing limited funds. They identify the relevant meaning before the original sentence is removed. This initial exposure is necessary because an unsupported demand to retrieve a completely unknown form would test guessing rather than retrieval. The example illustrates a design and does not report classroom data.

In the next phase, a learner recalls the meaning or supplies the missing target word in a new but comparable sentence. The system is instructed to request one attempt and then offer a choice of support. A cue might draw attention to the relationship between a resource and its intended recipient. If the learner cannot proceed, a direct explanation becomes available. After feedback, the learner writes a sentence about allocating time within a familiar project and explains the resource and purpose expressed. The teacher reviews disputed examples, particularly when an AI response accepts awkward phrasing or introduces a different sense of the word. Table 2 summarises the sequence and the evidence each stage is intended to produce.

Table 2-Illustrative prompt sequence for a teacher verified vocabulary item

Stage	Prompt or instruction	Learner evidence
Orient	Read the verified example and identify the resource and purpose	Initial interpretation

Attempt	Hide the example and recall the meaning or target form	Unaided retrieval attempt
Support	Choose a cue, an example, or a direct explanation	Support choice and response
Evaluate	Explain how the feedback changes your interpretation	Reasoned acceptance or challenge
Reuse	Write a new project sentence and justify the word choice	Contextualised production
Revisit	Recall and use the item in a later task without AI	Delayed independent performance

A reusable teacher instruction could read: “Use only the verified target meaning and examples supplied. Ask one question at a time. Invite an initial attempt, then offer a cue or explanation. Give specific feedback on the target word. Ask the learner to explain one revision and use the word in a new situation. If uncertain about correctness, flag the example for teacher review.” This instruction is a proposed scaffold, not a guarantee that a model will comply. Teachers should preview outputs, monitor departures from the target, and retain a verified fallback explanation. Prompt compliance and feedback accuracy are separate matters.

The same sequence can be implemented with printed cards, a shared device, or teacher prepared examples. Students can record an initial answer privately, select a hint card, discuss feedback with a partner, and complete a later individual retrieval activity. This equivalence preserves the intended learning opportunity where reliable individual access is unavailable. GenAI adds potentially flexible follow-up and varied examples, but those advantages must be weighed against verification work, connectivity, and unequal familiarity with prompting. A deliberately comparable offline option also creates a useful research contrast: it helps determine whether benefits arise from the learning sequence or from features specific to conversational AI.

The teacher can make evaluation concrete by contrasting two candidate sentences: “The committee allocated funds to the library” and “The committee allocated the problem yesterday.” Learners identify why the first expresses distribution of a resource and why the second does not demonstrate the target meaning. They then propose a different resource and purpose, such as allocating time to planning. These constructed examples illustrate the intended distinction and are not learner responses. A brief rubric can separate accurate meaning, appropriate grammatical use, and a relevant explanation. Scoring those dimensions separately prevents an articulate explanation from concealing an inaccurate sentence. Later assessment should introduce a different situation rather than repeat either example, allowing the researcher to distinguish recall of the model sentence from flexible use of the target.

Four propositions and opportunities for disconfirmation

The following propositions specify directional expectations derived from the synthesis. They concern bounded comparisons rather than guaranteed effects. Each assumes that target vocabulary, assessment demands, and opportunities for feedback are appropriate to the learners. The propositions should be tested separately before a bundled intervention is described as validating the full model. Table 3 connects each prediction with evidence that would weaken its proposed explanation.

Table 3-Testable propositions and disconfirming observations

Proposition	Predicted comparison	Observation challenging the explanation
P1 Retrieval	Attempt before feedback improves delayed recall over explanation before attempt	No recall advantage despite verified retrieval and comparable exposure
P2 Contingent support	Graduated support particularly benefits learners with weaker prior knowledge	Equivalent benefits across knowledge levels, or persistent overload
P3 Evaluative reuse	Justified contextual reuse improves novel use beyond copying examples	Gains confined to trained sentences or meaning recognition
P4 Learner control	Informed support choice improves regulation as scaffolds are reduced	Choices remain uncalibrated and independent performance declines

P1 predicts that an attempt before accurate corrective feedback will improve delayed recall relative to receiving the same explanation before attempting the item, when learners have previously encountered the target and practice opportunities are comparable. The mechanism is retrieval rather than a generic increase in time spent chatting. A convincing comparison would hold content, later feedback, and exposure opportunities as similar as practicable, while measuring actual time. An immediate test alone would be insufficient because the conditions differ in recent support. No delayed advantage, despite evidence that retrieval occurred, would weaken the prediction for the specified context. Any benefit would still require checking whether additional exposure or unplanned rehearsal supplied an alternative explanation.

P2 predicts an interaction between prior knowledge and the organisation of assistance. Learners with less relevant knowledge should benefit more from accessible cues and worked examples than from repeated unaided attempts, whereas learners with stronger knowledge may need less support. The proposal combines learning relevant challenge with limits imposed by available knowledge; it does not presume a universal optimal difficulty. Prior knowledge should be

measured separately from self confidence, and proficiency should not be inferred from institutional level alone. Evidence of unchanged overload, high abandonment, or no differential response to support would challenge the proposed matching mechanism. An apparent benefit limited to learners already skilled at prompting would identify an additional boundary.

P3 predicts that generating a new sentence and explaining a lexical choice will improve appropriate use in unfamiliar contexts relative to copying or lightly editing an AI example. The intended operation is evaluation of the relationship among meaning, collocation, and communicative purpose. Comparisons must account for the extra production and exposure that a justification task may introduce. They should include blinded assessment of novel contexts rather than reward similarity to the training sentence. An advantage confined to the practised wording would suggest memorisation or task specific rehearsal rather than broader productive knowledge. Similarly, improved explanations without improved use would indicate that explicit reasoning and productive performance have diverged.

P4 predicts that meaningful support choices, accompanied by reflection on their usefulness, will improve calibration of help seeking as prompts are gradually reduced. Compared with a fixed sequence, learners should increasingly select assistance that matches demonstrated need and maintain independent performance. Autonomy is therefore defined through informed regulation rather than low chatbot use or resistance to help. The prediction would be weakened if choices remain unrelated to need, if learners routinely request complete answers without subsequent processing, or if performance deteriorates when guidance is removed. Short term satisfaction would not resolve these possibilities. Repeated observations are necessary to distinguish developing regulation from novelty, compliance, or avoidance.

Research design and measurement priorities

An initial evaluation should establish feasibility before testing broad effectiveness claims. Researchers can pilot whether learners understand the instructions, whether prompts elicit the intended operations, and whether generated feedback remains accurate across plausible responses. A small pilot can identify ambiguous cues, excessive reading demands, and differences in access. It cannot establish a generalisable causal effect merely because participants complete the activity. The next stage should compare clearly specified sequences and include a delayed assessment aligned with the primary proposition. Interviews or stimulated recall can clarify learners' interpretations, but retrospective accounts should be considered alongside recorded activity and independently scored outcomes.

Treatment assignment must match the level at which instruction differs. If whole classes receive different conditions, sufficient classes are needed to distinguish the intervention from class and teacher characteristics. One class per condition cannot disentangle those sources of difference through individual level significance testing. Alternatively, learners or target items may be assigned within

a feasible design that addresses contamination and carryover. The choice should follow the educational setting and intended causal claim. Sample planning should consider clustering, repeated observations, expected precision, and attrition rather than adopt an arbitrary total. These are proposed design requirements; the present article supplies no estimate of an ACEM treatment effect.

Assessment should distinguish meaning recognition, meaning recall, form recall, and contextual use where the research question requires them. Test forms should avoid reproducing the exact AI generated examples, and scoring criteria for productive responses should specify acceptable meanings and uses. Independent raters can assess a subset of responses to establish agreement and resolve uncertainty. Pretesting can document prior knowledge but may itself create an encounter with the targets, so its role should be reported. Immediate and delayed tests should be scheduled in advance, and intervening practice should be documented. Claims should remain proportionate to the dimensions of knowledge and the retention intervals actually measured.

Process measurement should record the sequence of learner attempts, support choices, feedback, and revisions with appropriate consent. Logs can identify whether an answer was requested before an attempt, but cannot directly reveal covert retrieval or comprehension. Brief explanations, selected verbal reports, and carefully chosen self-reports can supplement behavioural traces. Load and engagement questionnaires require attention to wording, language, and construct validity. Analyses should not label the same measure as involvement, engagement, and cognitive load in different sections. When mediation is proposed, temporal ordering and the assumptions required for a causal interpretation must be explicit; a correlation between message counts and scores is insufficient.

A useful comparison could separate the presence of conversational AI from the presence of the learning sequence. One condition would use the staged sequence with a chatbot, another would use the same sequence with verified static materials, and a third would provide direct AI explanations. The first comparison addresses the additional contribution of contingent dialogue; the second examines the bundled sequence within an AI environment. Neither alone isolates every component. A subsequent focused experiment could manipulate only feedback timing or support choice. This progression makes competing explanations visible and avoids attributing an intervention package to a single attractive feature. All conditions should have credible educational value, and the chosen contrast should be stated before outcomes are analysed.

Researchers should also distinguish active task time from connection delays and time spent repairing misunderstood instructions. These durations can have different educational meanings even when they produce the same total session length. A sequence that generates more interaction because the model repeatedly misinterprets requests has not necessarily increased learning relevant effort. Conversely, a short exchange may follow substantial private retrieval or discussion with a partner. Combining event records with selective observation can clarify

these possibilities without assuming that continuous surveillance is necessary. Such evidence would help determine whether the proposed mechanism operated and whether the activity remains feasible under ordinary classroom constraints.

Feedback reliability should be assessed independently of prompt compliance. A system may follow every conversational instruction while accepting an incorrect meaning, and an accurate explanation may arrive too early for the intended retrieval opportunity. Researchers can therefore code these dimensions separately. When errors occur, they should examine whether learners notice them, whether teachers repair them, and whether the mistaken form or meaning appears later. This approach treats unreliable feedback as a possible moderator of subsequent learning, rather than quietly removing problematic exchanges from the analysis of instructional effectiveness.

Transparent reporting must document the model or service used, available version information, dates of use, prompts, relevant settings, interaction constraints, and procedures for checking feedback. When a commercial system does not disclose a stable version or a setting, this absence should be recorded rather than filled with an assumption. Researchers should report departures from the intended sequence, exclusions, missing observations, and access interruptions. Estimates with uncertainty, descriptives, and appropriate comparison outcomes are more informative than isolated significance statements. Anonymised materials and feasible examples of interaction traces can improve interpretability, while raw conversations should not be shared automatically when they contain identifiable or sensitive information.

Teacher agency digital equity and responsible implementation

ACEM makes teacher judgement an instructional component. Teachers select useful vocabulary, judge whether a contextual distinction matters, and determine when an explanation should replace further attempts. This role responds to the concern that educational AI research can give insufficient attention to educators and pedagogy (Zawacki-Richter et al., 2019). It also accords with Kohnke et al.'s (2023b) account of language instructors' needs for AI related competence and institutional support. Training should include evaluating alternative feedback, identifying inaccurate examples, and adapting sequences to learning goals. Mastering elaborate prompt formulas is a poor substitute for understanding the lexical decision that the lesson asks learners to make.

AI literacy should be integrated into ordinary learning decisions. Ng et al. (2021) connect understanding and using AI with evaluation and ethical concerns. In a vocabulary activity, learners can compare an AI explanation with a teacher approved resource, identify an unsupported usage claim, and explain why an example fits a particular context. These actions make evaluation consequential for the language task. They should not become a separate technical obstacle for students with limited digital experience. A short modelled comparison, a visible support menu, and familiar instructional language may provide a more accessible entry than requiring each learner to construct a complex prompt independently.

Teacher workload deserves explicit evaluation. Preparing accurate examples, checking unusual responses, and reviewing learner explanations may offset time saved through automated generation. A feasible implementation can begin with a limited vocabulary set and a shared bank of checked feedback, expanding only when the teacher can maintain quality. Studies should document preparation and supervision demands alongside learner outcomes. This information would help institutions decide whether any measured advantage is sustainable and whether teachers need additional time, resources, or professional support to implement the sequence consistently.

Equitable implementation requires analysing the costs of the activity rather than assuming universal access from device ownership. Data charges, unreliable connectivity, shared devices, and account restrictions can change who receives feedback and how often practice occurs. In resource constrained EFL settings, the teacher can prepare verified examples centrally and provide equivalent retrieval and feedback opportunities through print or shared classroom interaction. Research should record access conditions and avoid interpreting missing interaction as low motivation. Accommodations may include additional processing time, readable text, oral response options, and explanations in a familiar language. These adjustments preserve the target learning operation while reducing demands unrelated to vocabulary knowledge.

Ethical considerations are likewise part of implementation quality. Akgun and Greenhow (2022) identify privacy, bias, and related challenges in educational AI, particularly in school settings. An ACEM activity should minimise personal information in prompts and clarify who can view retained conversations. Learners need a route to question feedback and seek human assistance. Examples should be reviewed for stereotyping, cultural assumptions, and misleading claims about acceptable English. The model does not require commercial AI use as a condition of participation. Where institutional rules or learner needs preclude such use, equivalent tasks can preserve access to the educational objective.

Limitations and theoretical development

The conceptual synthesis is selective and cannot establish that ACEM improves engagement, vocabulary retention, or autonomy. Its integration of several traditions risks excessive breadth unless future studies distinguish the constructs and test a limited set of relationships at a time. The proposed sequence may also overlap with well established teaching practices. That overlap is deliberate at the level of learning principles, but the model warrants continued use only if specifying the division of work in AI dialogue adds explanatory or practical value. Comparative studies should therefore include credible instructional alternatives and report situations in which the additional technology offers little benefit.

Further limits concern learners, language, and technological change. The illustrative activity assumes some relevant vocabulary knowledge and capacity to interpret feedback. Younger learners, beginners, learners with disabilities, and users working across several languages may require different forms of support.

Productive use also involves pronunciation, register, and communicative appropriateness that a short written task cannot represent fully. Model updates can alter feedback quality and instruction following, limiting reproducibility across time. The most defensible research programme will examine these boundaries directly, retain negative findings, and revise the framework when its predictions fail. Generalisation should follow converging evidence across contexts rather than the popularity of a tool.

IV. CONCLUSION

GenAI assisted vocabulary learning should be evaluated through what learners can understand and use after assistance is removed. ACEM proposes that prompt sequences can protect opportunities to attempt, evaluate, revise, and retrieve, provided that the demands remain accessible and feedback is dependable. Productive friction is consequently a bounded instructional proposal whose value requires evidence, rather than a justification for making learning slower or more frustrating. The framework separates design features, engagement, perceived burden, and lexical outcomes so that these relationships can be investigated without conflating their measures. The practical implication is to begin with the target knowledge and assign each part of the work deliberately. Teachers can provide initial exposure, invite an attempt, offer proportionate help, require meaningful reuse, and return to the item after an interval. Learners retain a role in selecting support and questioning feedback, while equivalent activities preserve participation across access conditions. The four propositions provide a research agenda for testing retrieval, contingent support, evaluative reuse, and developing regulation. Their contribution depends on transparent comparisons and delayed independent assessment. Evidence that challenges the model will be as useful as evidence supporting it in defining when conversational AI serves EFL vocabulary learning.

REFERENCES

- Aziz, I. N. (2024). Leveraging Memory Strategies to Enhance Vocabulary Proficiency In English Language Education At An English Course In Indonesia. *JEET, Journal of English Education and Technology*, 5(03), 190–205. <https://doi.org/10.59689/jeet.v5i03.145>
- Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2, 431–440. <https://doi.org/10.1007/s43681-021-00096-7>
- Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, 1, 496–511. <https://doi.org/10.1038/s44159-022-00089-1>

- Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2024). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research*, 28(1), 201–230. <https://doi.org/10.1177/13621688211001289>
- Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23, 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Karataş, F., Abedi, F. Y., Gunyel, F. O., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 29, 19343–19366. <https://doi.org/10.1007/s10639-024-12574-6>
- Karatas, N. B., Özemir, O., Lovelett, J. T., Demir, B., Erkol, K., Verissimo, J., Erçetin, G., & Ullman, M. T. (2025). Improving second language vocabulary learning and retention by leveraging memory enhancement techniques: A multidomain pedagogical approach. *Language Teaching Research*, 29(1), 112–149. <https://doi.org/10.1177/13621688211053525>
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. <https://doi.org/10.1126/science.1152408>
- Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, Article 100225. <https://doi.org/10.1016/j.caeai.2024.100225>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023a). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. <https://doi.org/10.1177/00336882231162868>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023b). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence*, 5, Article 100156. <https://doi.org/10.1016/j.caeai.2023.100156>
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>

- Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45, 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence*, 7, Article 100266. <https://doi.org/10.1016/j.caeai.2024.100266>
- Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, 9, Article 100445. <https://doi.org/10.1016/j.caeai.2025.100445>
- Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach. *Computers and Education: Artificial Intelligence*, 7, Article 100278. <https://doi.org/10.1016/j.caeai.2024.100278>
- Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, 37(2), 233–260. <https://doi.org/10.1177/0267658320927764>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, Article 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, Article 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, 36, 50–72. <https://doi.org/10.1017/S0267190515000094>
- Pyke, W., Lunau, J., & Javadi, A.-H. (2025). Does difficulty moderate learning? A comparative analysis of the desirable difficulties framework and cognitive load theory. *Quarterly Journal of Experimental Psychology*, 78(10), 2181–2195. <https://doi.org/10.1177/17470218241308143>
- Reeve, J., & Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology*, 36(4), 257–267. <https://doi.org/10.1016/j.cedpsych.2011.05.002>

- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 53(1), 109–120. <https://doi.org/10.1017/S0261444819000326>
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. <https://doi.org/10.1177/1745691615569000>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31, 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69(3), 559–599. <https://doi.org/10.1111/lang.12343>
- Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33–52. <https://doi.org/10.1017/S0272263105050023>
- Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, 71(2), 487–536. <https://doi.org/10.1111/lang.12444>
- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 437, pp. 1–21). Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581388>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—The educators have been forgotten? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>