Generative AI tools in designing MCQs for English language examinations: Insights from Lecturers

Vu Thi Kim Chi[1], Nguyen Trinh To Anh[1*], Vo Dao Vuong Co[1]

[1] Saigon University, Hochiminh City, Vietnam
[*]Corresponding author's email: nttanh@sgu.edu.vn
[*] https://orcid.org/0000-0003-4574-1681
https://doi.org/10.54855/979-8-9870112-9-4_1

® Copyright (c) 2025 Vu Thi Kim Chi, Nguyen Trinh To Anh, Vo Dao Vuong Co

Received: 01/09/2025 Revision: 03/12/2025 Accepted: 08/12/2025 Online: 17/12/2025

ABSTRACT

This study examines ChatGPT's ability to generate syntax-based sentence transformation multiple-choice questions (MCQs) using the syntactic types listed in the ReParaphrased classification framework. These include: Negation Switching (NS), Diathesis Alternation (DA), Subordination and Nesting Changes (SNC), Coordination Changes (CC), and Ellipsis (Ell). Using a quantitative approach, the researchers aim to provide personal insights into designing exam questions based on the content of B1 Empower. A statistical analysis of 120 AI-generated test items was conducted to identify the frequency and distribution of each syntactic transformation type, highlighting the favored patterns in the generated dataset. The findings suggest that ChatGPT tended to create test items using tactics with a clear pattern of transformation, such as SNC and NS, while showing less favor for tactics that require more nuanced contextual understanding, such as Ell and CC. In addition, AI could create questions quickly and effectively; however, some problems remained, including semantic distortions and awkward forms, such as double negatives or passives. This result highlights the crucial role of human intervention in proofreading and refining AI-generated questions to ensure the accuracy and relevance of the dataset items.

Keywords: Generative AI tools, MCQs, AI-generated exam questions, lecturers' insights, syntax-based sentence transformations, ReParaphrased classification framework

Introduction

Generative AI offers a promising solution for English language evaluation by reducing educators' workload in creating multiple-choice questions (MCQs). Although MCQs enable quick and objective assessment across diverse language areas, designing effective items with well-crafted stems, correct answers, and plausible distractors remains challenging. Generally speaking, the stem must be legitimate while the distractors should be persuasive enough, plausible but incorrect (Dhawaleswar Rao & Saha, 2020).

MCQs can come in many forms, including check-box, matching, gapped-text filling, and sentence transformation. In a sentence transformation task, students are tested on their ability

CITATION | Vu, T. K. C., Nguyen, T. T. A., & Vo, D. V. C. (2025). Generative AI tools in designing MCQs for English language examinations: Insights from Lecturers. *ICTE Conference Proceedings, 9*, 1-19. ISSN: 2834-0000, ISBN: 979-8-9870112-9-4. DOI: https://doi.org/10.54855/979-8-9870112-9-4_1

to recognize different ways to express the same idea, a useful skill not only in academic settings but also in daily communication. In this question form, sentence structure is modified using different techniques without affecting the meaning. For instance, direct speech can be transformed into indirect speech, or passive voice can be converted into active voice. Additionally, different vocabulary can be used to convey the same idea. Mastering this skill proves students' versatility in language usage.

When designed as MCQs, sentence transformation tasks assess whether students can recognize accurate paraphrases while avoiding distractors that distort meaning. For instance:

Original sentence: Although he was tired, he finished the report. *(Stem)*

Question: Which sentence has the same meaning?

A. He finished the report because he was tired. *(Distractor)*

B. Despite being tired, he finished the report. *(Correct answer - Key)*

C. He finished the report, so he was tired. *(Distractor)*

D. Because he finished the report, he was tired. *(Distractor)*

Compared to essays or short-answer questions, their regulated format ensures higher dependability by minimizing subjectivity. MCQs are particularly helpful for extensive tests, as they offer practical benefits such as effective scoring and interoperability with automated systems. However, manually developing them requires significant time, resources, and mental energy. This load may cause quality to suffer, leading to questions that assess only rudimentary memory or contain pedagogical errors. Education has increasingly relied on technology to overcome the significant challenges of manually creating questions. The original purpose of Automatic Question Generation (AQG) was to meet the demand for a continuous flow of fresh, high-quality questions while reducing the costs and time required for manual production (Maas et al., 2024). A "new frontier" in this field is the rise of large language models (LLMs), such as ChatGPT (Isley et al., 2025). These tools provide quick, dependable assistance in creating diverse, standards-aligned, and differentiated questions for all learning levels. They can generate unique material from a wide range of instructional patterns (The Case HQ, 2025).

With the potential to automate question generation, reduce administrative burden, and free up teachers to focus on more meaningful educational exchanges, the emergence of Generative Artificial Intelligence (GenAI) presents an appealing option. This is not a cure-all, however; using AI raises a number of issues of its own, such as errors, algorithmic bias, and reduced higher-order thinking.

This study explores ChatGPT's ability to generate MCQs via syntax-based sentence transformations, following the ReParaphrased classification framework developed by Zhou. et al. (2025). Analysis focused on five syntactic types: Negation Switching (NS), Diathesis Alternation (DA), Subordination and Nesting Changes (SNC), Coordination Changes (CC), and Ellipsis (Ell).

Literature review

The Linguistic and Typological Foundations of Paraphrase

Core Linguistic Theories: Chomsky and Generative Semantics

Language is not just a set of rules, but also a special *human* intellectual ability. Chomsky argues:

"Current approaches cannot capture the creative nature and generative capacity of human language." (Chomsky, 2006). With the *publication of* Syntactic Structures (1957), Chomsky broke *with* the traditional framework, bringing about a revolution in *how* linguistics is perceived. This work laid the foundation for generative grammar theory (i.e., the First Linguistic Model), opening a new era in language research that focused on the innate ability and creativity of human language. The generative grammar incorporated three aspects of language: syntax, phonology, and semantics. Chomsky revolutionized the study of language with his Transformational-Generative Grammar model (i.e., the Standard Theory and the Extended Standard Theory), published in Aspects of the Theory of Syntax (1965), which analyzes the structure of language at both the abstract and *the* concrete levels. This model consists of two basic structures: Deep Structure, *which* reflects the *sentence's* abstract meaning *and is* considered universal across languages, and Surface Structure, *which is* the way the sentence is expressed *and* varies *by* language. For instance, the transformati*on* between active voice, such as "The cat ate the fish," and passive voice, such as "The cat ate the fish," *is* frequently used to convey the same meaning while serving different communicative purposes. Chomsky's transformational-generative grammar provided a rule-based framework for understanding how humans produce complex, creative sentences and the relationship between form and meaning. His theories, including the formal language taxonomy, have significantly influenced modern linguistics and related fields such as NLP, AI, and automation.

In the late 1960s and early 1970s, Generative Semantics, led by scholars such as Lakoff, Ross, McCawley, and Postal, expanded on Chomsky's work by proposing that deep structure represents meaning itself rather than merely syntax (McCawley, 1968; Postal, 1974; Lakoff & Ross, 1976). This perspective held that at the most basic level of analysis, syntax and semantics were inseparable. This theory's main supporting evidence was paraphrase, which holds that two different surface structures must inevitably have the same underlying deep (semantic) representation if they express the same meaning. For example, because they both came from a single, common semantic representation of a "transfer" event, the sentences "John gave Mary a book" and "John gave a book to Mary" were regarded as paraphrases. *Generative Semantics questioned the* rigorous division between syntax and semantics, sparking an important discussion in the *field* that influenced later ideas.

Chomsky's framework laid the theoretical groundwork for paraphrase in the mid-20th century by distinguishing between a sentence's underlying deep structure and its observed form, which provided the fundamental explanation for why two different surface forms could express the same underlying proposition. Classic transformational operations, such as active-passive voice alternation, negation, and ellipsis, served as the formal syntactic mechanisms for mapping a single deep structure to multiple surface forms. The Generative Semantics, while challenging Chomsky's model, reinforced the idea that paraphrases are sentences that ultimately share the same underlying semantic deep structure, even if they are syntactically distinct, and that paraphrase works as a crucial event for comprehending the relationship between meaning and form in language.

Modern Typologies of Paraphrasing

Modern language testing and instruction draw heavily on transformational grammar and computational linguistics, emphasizing tasks that transform sentence structure while preserving meaning. Research supports the pedagogical value of different transformation types, such as Negation Switching (NS), which tests learners' ability to retain propositional meaning (Vahtola et al., 2022), and Diathesis Alternation (DA), which evaluates understanding of voice and argument structure (Thompson et al., 2013). Subordination & Nesting Changes (SNC) and

Coordination Changes (CC) assess learners' skill in modifying clause structures without altering meaning, providing clear criteria for test design (Vila et al., 2014). Finally, Ellipsis (ELL) tasks assess comprehension by examining how readers infer missing information from sentence structure, supported by studies showing that readers can infer missing information from sentence structure (Kim et al., 2020).

Paraphrase in Applied Linguistics: Pedagogy and Assessment

From Structural Drills to Communicative Practice

In the context of audiolingual and structuralist classroom instruction, the language theories of the day had a significant and direct influence on pedagogical practices. Transformation drills, a fundamental classroom practice, were theoretically justified by the idea of transformational processes, which demonstrated the connection between deep and surface structures. Classroom activities were designed *to help students* learn to modify language patterns, such as *switching between the* active and passive voice. These drills were implemented in the audiolingual classrooms in the 1950s, following *Fries's (1945)* promotion of such pattern practices as early as the 1940s. Later, Lado (1964) scientifically justified the use of such drills as crucial instruments *for mastering sentence-structural* changes.

As language teaching approaches real-life communication, the use of transformation practices changes. Instead of repetitive drills, teachers began to focus on meaningful paraphrasing activities. Widdowson (1978) noted that conventional transformation drills in audiolingualism were so academic that they could not help learners use language in authentic communicative contexts. He suggested that these *changes in sentence structure* could improve learners' communication skills. Nunan (1989) built on this idea, demonstrating how structural drills could be transformed into communicative tasks *in which* paraphrasing and rewording played a key role in enhancing learners' versatility *in* language skills. Similarly, Nation (2009) stressed that paraphrasing and rewording are not just about changing sentence structure – they are essential skills for communicating clearly in academic and professional settings, connecting old structural methods with contemporary focus-on-meaning contexts.

Paraphrasing as a Core Academic and Communicative Skill

Based on these ideas, modern applied linguistics considers paraphrasing a crucial skill for effective communication rather than just a grammar-focused task. Certain research was conducted at tertiary institutions. Na and Mai (2017) uncovered students' perspectives on the importance of paraphrasing in their coursework and found that participants lacked knowledge of paraphrasing techniques, particularly regarding changes in syntactic structures. Tran and Nguyen (2022) found that paraphrasing positively impacts EFL students' academic writing performance. These studies show that good paraphrasing does not merely replace words; it also involves reordering the particles, entities, and verbs of a sentence to convey the same meaning as the original sentence (Hosking & Lapata, 2021; Zhou et al., 2019). This explains why paraphrasing should not be considered a routine activity but a powerful tool for knowledge transformation and persuasive communication (Hirvela & Du, 2013). The innovative view of paraphrasing in applied linguistics emphasizes its role in facilitating flexible, accurate communication (Chen et al., 2015).

Assessing Sentence Transformation in ESL/EFL Testing

Language tests and textbooks *began* to include *elements of* sentence transformation for both form-focused and meaning-focused objectives, reflecting this pedagogical shift (i.e., communicative teaching). These materials *began* incorporating exercises *in which* learners had to rewrite sentences, which helped *assess* both their knowledge of grammar and their understanding of meaning. Early books on language testing (Heaton, 1979) showed how these exercises could test *students'* grammar *skills*. Although fellow writers, such as Alderson et al. (1995), criticized the tasks' shortcomings in evaluating genuine communication skills, they were still useful for checking correct grammar and matching meanings. Later research (Paribakht, 2004), which emphasized that tasks such as paraphrasing increase semantic processing, made the dual purpose of these activities clear. In his 2003 textbook, Anthony Hughes noted that transformation exercises continued to be used, but they evolved into a more adaptable instrument for evaluating both conventional grammatical knowledge and the contemporary, meaning-based ability *to paraphrase*.

Sentence transformation tasks now play a key role in modern standardized language tests. These tasks are used to assess both how well students can correctly modify sentence structures and whether they retain the original meaning. Heaton (1979) and Alderson et al. (1995) suggested that the tasks evaluate students' ability to control grammar, marking an early investigation in writing testing. Later, Hughes (2003) revealed the communicative purposes of such tasks. The exercise of expressing the same idea in different ways has long *been a* well-known feature of language testing examinations. Test-takers can encounter the so-called task "keyword transformations" in the Use of English paper of Cambridge English Exams. The task requires them to use the given *keyword and apply* their knowledge of structural sentences to complete the paraphrased sentences. Hamilton's (2025) research on this task supports insights into the importance of *a* variety of sentence structures in meaning delivery, although it has long been a *major* topic of discussion in the history of stylistics. In other words, sentence transformation tasks remain important for assessing both grammatical knowledge and the ability to rephrase ideas in various ways.

The Automation of Paraphrase: AI in Language Generation and Evaluation

AI and Paraphrase Generation

In the world of computational linguistics, paraphrase identification (PI) and its associated database have made significant contributions to the development of Natural Language Processing (NLP) (Zhou et al., 2025). AI systems like ChatGPT and Gemini can automatically transform large numbers of sentences by using extensive real-life examples to generate paraphrases with high syntactic and semantic accuracy. Research on paraphrasing typologies further supports this process by classifying transformations at the lexical, semantic, and structural levels, thereby improving AI model performance (Wahle et al., 2023).

Automated Evaluation of Paraphrase and MCQ Generation

Additionally, a variety of automated metrics have been created to assess whether AI-generated changes maintain the original meaning. Even in cases *with* substantial surface variance, embedding-based metrics such as the Universal Sentence Encoder (USE) and BERTScore provide reliable, automated methods for evaluating semantic *equivalence*. These technologies are essential for automatically evaluating the quality of MCQs *generated* by AI and can *serve* as a scalable alternative to human assessment. Transformer models are potential and current best practices in this field. Rodriguez-Torrealba et al. (2022) studied the efficiency of neural language models in generating multiple-choice question*naire*s. The results, assessed by experts, showed that the questionnaires contained both well-formed question stems and realistic distractors. Mulla and Gharpure (2023) established methods to evaluate whether AI can successfully replicate the pedagogical and linguistic goals of transformation-based tasks.

About this Research

In this paper, by evaluating AI-generated sentence transformation MCQs, we test whether machines can replicate what transformational grammar aimed to capture, the recognition of different forms that share an underlying proposition.

The concept of paraphrase has its theoretical roots in transformational grammar, where different surface realizations are understood to share the same underlying proposition (Chomsky, 1957). This theoretical foundation later shaped pedagogical practices in ESL/EFL, where sentence transformation exercises have been widely adopted to assess learners' ability to recognize and produce equivalent meanings across varied syntactic forms (e.g., active–passive alternation, negation, subordination). The authors also acknowledge that there are certain restrictions on students' ability to paraphrase in academic settings. Na and Mai (2017) examined students' paraphrasing abilities and *found* challenges *with* syntactic changes, as they primarily employed synonym replacement. This research aims to provide a *range* of paraphrasing typologies *to* assist teachers in language testing *and* learners in developing paraphrasing literacy.

Recent paraphrase research has been organized into typological frameworks, such as the Extended Paraphrase Typology (EPT) by Kovatchev et al. (2018), which categorizes paraphrases at the lexical, syntactic, and semantic levels. Thereby, Zhou et al. (2025) introduced ReParaphrased, a more detailed framework that identifies structure-based changes - such as Diathesis Alternation (DA), Negation Switching (NS), Ellipsis (ELL), Subordination and Nesting Changes (SNC), and Coordination Changes (CC) - for improved paraphrase analysis and computational modeling.

Table 1.

The syntax-based types of paraphrases in ReParaphrased

Classification	**Explanation**	**Original Sentence**	**Paraphrased Sentence**
Negation Switching (NS)	Replacing a word with its antonym and applying a negation	They weren't unharmed.	They were harmed.
Diathesis Alternation (DA)	Reorganizing entities and actions by switching between different syntactic voices or argument structures of a verb without changing the underlying meaning	His mother bought him a present.	He was bought a present by his mother.
Subordination & Nesting Changes (SNC)	Altering the hierarchical structure of a sentence, such as turning a main clause into a subordinate clause or embedding information inside a relative clause	Because she was late, she missed the bus.	She missed the bus, because she was late.
Coordination Changes (CC)	Altering how ideas are linked, such as switching between coordinate conjunctions or reordering coordinated elements while maintaining meaning	He is smart and this helps him learn easily.	Because he is smart, he can learn easily.
Ellipsis (Ell)	Omission of words or phrases that are understood from the context, resulting in a shorter but semantically equivalent sentence	She went to the store, and she bought some milk.	She went to the store to buy some milk.

This study applies the five ReParaphrased syntactic categories to analyze AI-generated sentence-transformation questions, aligning with transformational-grammar principles commonly used in language teaching and computational-linguistics research. Our objective is to verify if machine-generated items accurately reflect the types of paraphrase transformations that educators and test developers have used in classroom assessments for a long time.

Research Question

The study aims to examine how syntactic transformations were distributed across the five ReParaphrased categories. To fulfill this aim, the study seeks to answer the research questions:

1. How syntactic patterns were distributed across the five ReParaphrased categories in four-option generated MCQs by ChatGPT?

2. What are lecturers' insights on ChatGPT's assistance in MCQ design regarding syntactic transformations?

Methods

This study employed a mixed-methods approach to examine the capabilities of large language models (LLMs) for generating syntax-based multiple-choice questions for English-language assessment. The method involves using a specific LLM model to create a corpus of 120 multiple-choice questions with a controlled prompt, followed by a two-phase analysis: a qualitative analysis to classify the questions according to the REParaphrased framework, and a statistical analysis to identify the generation patterns. This designated approach aimed to rigorously examine the LLM's application in assisting teachers in their assessment.

Data corpus

This corpus consisted of 120 four-option MCQs to assess grammar and vocabulary competence. This selection of 120 questions is a comprehensive, diverse, and representative sample for empirical analysis and the identification of clear generation patterns, yet manageable for detailed manual qualitative analysis.

The questions were designed to assess the linguistic competence of non-English-major university students at the B1 proficiency level. At this level, learners are required to have a grasp of common grammatical structures and a moderate volume of vocabulary, making it appropriate for assessing their language competence. This is also the level of language proficiency required of non-English-major Vietnamese students to graduate from university.

These questions were designed based on the grammar and vocabulary content of Units 1-6 of Empower B1. With a broad coverage of B1-level grammar and vocabulary, this book ensured that the generated questions were contextualized, relevant to typical B1 curricula, and consistent in difficulty and thematic content. Furthermore, its widespread use in many university English language programs in Vietnam enhances the research's practicality, making it applicable to the current teaching context.

All the items in the set of questions were designed in a syntax-based sentence transformation format, requiring students to recognize and identify paraphrased sentences that use different structures. The classification of these transformations followed the REParaphrased framework, which provides a comprehensive system for classifying syntactic changes. This framework was chosen for its ability to offer insights into how sentences can be manipulated structurally, which is crucial for analyzing GPT-5's generation patterns.

Data generation process

GPT-5 was selected for its accessibility, ease of use, and efficiency in generating a large number of questions. To ensure corpus quality, the prompt engineering was carefully designed and thoroughly reviewed before being used.

First, the LLM was provided with definitions and illustrative examples of the five syntactic transformation types as defined in the REParaphrased framework. This was intended to help GPT-5 fully grasp the framework used to paraphrase the stem, ensuring the accuracy of the generated question items.

Then, the concept-checking process was conducted using trial materials to determine if the platform could generate questions accordingly. A pilot study was conducted before the main data generation process to fine-tune the LLM's understanding and ensure its ability to generate questions that met the study's specific requirements. Three preliminary content samples, distinct from the six units chosen but representative of B1-level grammar and vocabulary, were used as input to test the platform's ability to generate the correct types of questions within the given framework. During the pilot phase, the question items were reviewed to ensure they were syntactically correct and precisely aligned with the REParaphrased categories, or that they did not contain distractors that were too complex for B1 learners. The prompts were then adjusted several times regarding sentence complexity, distractor nature, and syntactic shifts within the framework.

Once all misunderstandings were settled, the full prompt was created, including detailed instructions on the number of options, the level of the test items, and the content on which they were based. To be more specific, each testing item included four options at level B1, aligning with the content of the Empower B1 units.

Data analysis

The analysis of the generated question items was conducted in two phases: a qualitative content analysis followed by a quantitative statistical analysis.

In the first phase, the corpus of 120 question items was manually reviewed and analyzed. This procedure was conducted independently by two researchers with a strong background in applied linguistics and extensive experience in English language teaching and testing. To ensure the reliability of the process, a training session was organized, which allowed the researchers to share their understanding of the REParaphrased framework and clarify any misunderstandings of its concepts. This is necessary to ensure consistent item classification and avoid miscategorizations that may occur as GPT-5 generates the dataset. After addressing the technical issues associated with the transformation, all items were proofread a second time to examine if there were any problems with language use. All the correct answers were checked for consistency in meaning and structural transformation to identify flaws in each syntactic transformation type. Consequently, each item's mistake was coded and linked to the type of syntactic transformation. This aimed to investigate whether those mistakes were repeated in other items that employed the same transformation technique.

In the second phase, a quantitative statistical analysis was performed by using Microsoft Excel. As the statistical analysis is relatively simple, there is no need to complicate the process by employing more advanced tools. In this stage, the frequencies and percentages of each of the five syntactic transformations (NS, DA, SNC, CC, Ell) were calculated to investigate their occurrence within the generated items. This aimed to highlight the most and least dominant types when GPT-5 generates questions of this type.

Findings and Discussion

Research question 1: How syntactic patterns were distributed across the five ReParaphrased categories in four-option generated MCQs by ChatGPT?

The analysis of the 120 AI-generated multiple-choice questions (MCQs) revealed distinct patterns in the distribution of syntactic transformations across the five categories defined by the ReParaphrased framework. For the first attempt, the proportions are visualized in the pie chart and summarized as follows:

Subordination & Nesting Changes (SNC): 34%

Negation Switching (NS): 27%

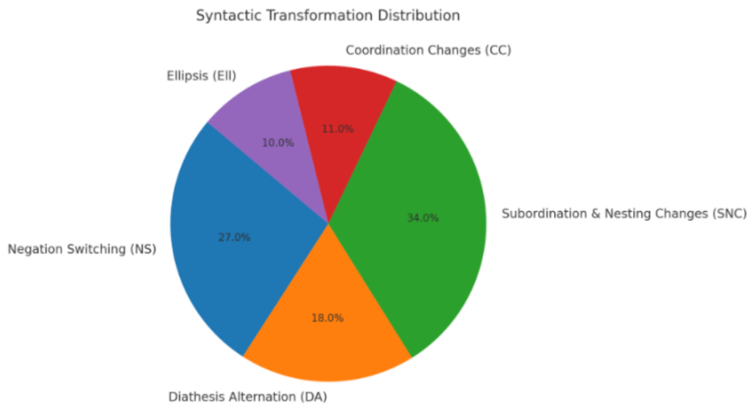
Diathesis Alternation (DA): 18%

Coordination Changes (CC): 11%

Ellipsis (Ell): 10%

The pie chart in Figure 1 shows the distribution of the five syntactic transformation categories, with Subordination & Nesting Changes (34%) being the largest and Ellipsis (10%) the smallest.

Figure 1.
Syntactic Transformation Distribution (1st attempt)



Subordination & Nesting Changes (SNC)

This was the most frequent category, comprising over one-third of the dataset. The AI often rewrote sentences by turning the main clause into a subordinate one or by shifting the way clauses were nested. Introducing subordinate clauses helps express *cause, time, contrast, or condition*, improving logical relationships between ideas. The AI favors these structures because they make sentences sound more connected and contextually rich. Overall, these items were mostly correct, but the model tended to rely too heavily on just a few conjunctions - especially *because, when,* and *while*.

Table 2.

Generated sentences with Subordination & Nesting Changes (SNC)

No.	Original sentence	Generated sentence (correct answer)
1.1.	We arrived at the airport late, *so* we missed the flight.	We missed the flight *because* we arrived at the airport late.
1.2.	*While* we were driving to the campsite, it started to rain.	We were driving to the campsite *when* it started to rain.
1.3.	He forgot to pack his passport, *so* he couldn't go abroad.	*Because* he forgot his passport, he couldn't go abroad.
1.4.	I was reading the guidebook *when* my friend came in.	*While* I was reading the guidebook, my friend came in.
1.5.	We exchanged some money *before* buying souvenirs.	We exchanged money *and then* bought souvenirs.
1.6.	She was unpacking her luggage *when* her phone rang.	Her phone rang *while* she was unpacking her luggage.
1.7.	*When* we were walking around the city, we got lost.	We got lost *while* we were walking around the city.
1.8	He just missed the bus, so he was late.	Because he missed the bus a short time ago, he was late.
1.9	There was a strike, *so* the trains didn't run.	The trains didn't run *because* there was a strike.
1.10	We had an adventure *when* we traveled around Asia.	We had an adventure *while* traveling around Asia.
1.11	I'm going to the hairdresser's *because* I want to look my best.	*Because* I want to look my best, I'm going to the hairdresser's.
1.12	He is looking tired, *so* he will go to bed early.	*Because* he is looking tired, he will go to bed early.

Negation Switching (NS)

Negation transformations accounted for just over a quarter of the dataset (27%) and were generally an area where the AI performed well. Most of the time, it can smoothly turn positive sentences into negative ones while keeping the meaning intact. However, a subset of items included awkward double negatives or forms that introduced slight semantic distortions, for instance, as seen in sentences 2.5, 2.8, and 2.10.

Table 3.

Generated sentences with Negation Switching (NS)

No.	Original sentence	Generated sentence (correct answer)
2.1	She *hardly ever* sends emails these days.	She *almost never* sends emails these days.
2.2	She *rarely* watches TV.	It's *unusual* for her to watch TV.
2.3	She *rarely* speaks in class.	She *almost never* speaks in class.
2.4	There is *something wrong* with the car engine.	The car engine *doesn't work properly*.
2.5	She *doesn't owe* me *any money*.	She *owes* me *no money*.
2.6	She *is not doing well* at school.	She *is failing to do well* at school.
2.7	I *haven't paid* her back yet.	I *still owe* her money.
2.8	She is going to wear boots, *not sandals.*	She *isn't going to* wear sandals, but boots.
2.9	She *isn't* wearing earrings.	She *doesn't have* earrings *on*.
2.10	*Don't* worry about the exam.	Worrying about the exam *isn't necessary.*

Diathesis Alternation (DA)

Diathesis alternation made up 18% of the dataset and was usually expressed through switching between active and passive voice. Most of these transformations were accurate and useful for teaching purposes, but they often felt repetitive, relying heavily on the same agent-patient structures. Moreover, some passive sentences require slight adjustments to sound more natural (3.5, 3.8 and 3.9).

Table 4.

Generated sentences with Diathesis Alternation (DA)

No.	Original sentence	Generated sentence (correct answer)
3.1	She's worried about her exams.	Her exams worry her.
3.2	You should discuss this problem with your teacher.	This problem should be discussed with your teacher.
3.3	He is interested in learning Spanish.	Learning Spanish interests him.
3.4	They advised me to take a rest.	I was advised to take a rest.
3.5	She suggested going to the park.	Going to the park was suggested *by her.*
3.6	She advised me not to skip breakfast.	I was advised not to skip breakfast.
3.7	He apologized for being rude.	His apology was for being rude.
3.8	You should save some money for the trip.	Some money should be saved for the trip *by you.*
3.9	She lent me her book.	Her book was lent to me *by her.*
3.10	They gave me a tip at the restaurant.	I was given a tip at the restaurant.

Coordination Changes (CC)

Coordination changes made up about 11% of the dataset, making them less frequent overall. In these cases, the AI typically modified how ideas were connected by switching between conjunctions like '*and*', '*or*', and '*but*', or by reordering coordinated parts of a sentence while maintaining the same meaning. The reason this type of transformation is least favored is that coordination changes often improve rhythm or emphasis rather than meaning, so the AI might not view them as essential edits.

Table 5.

Generated sentences with Coordination Changes (CC)

No.	Original sentence	Generated sentence (correct answer)
4.1	She spent money on new clothes *and* on shoes.	She spent money on new clothes *as well as* shoes.
4.2	He gave me a hug, and he gave me a smile.	He gave me *both* a hug *and* a smile.
4.3	We're going to the beautician, and we're going to the hairdresser's.	We're going to *both* the beautician *and* the hairdresser's.
4.4	This ancient building is magnificent, and it is peaceful.	This ancient building is *both* magnificent *and* peaceful.
4.5	The joke was silly.	The joke was stupid *and* not serious.
4.6	She bought sunglasses and a map.	She bought *both* sunglasses *and* a map.
4.7	She is wearing a jumper, and she is wearing tights.	She is wearing *both* a jumper *and* tights.
4.8	We exchanged some money before buying souvenirs.	We exchanged money *and then* bought souvenirs.

Ellipsis (Ell)

Ellipsis was the least common category, making up just 10% of the dataset. Generating options with this kind of transformation seemed difficult for ChatGPT because it required significant changes in grammatical structures, which could affect comprehension and demand extensive correction.

Table 6.

Generated sentences with Ellipsis (Ell)

No.	Original sentence	Generated sentence (correct answer)
5.1	We are going to get a new outfit, and we are going to have a shave.	We are going to get a new outfit and have a shave.
5.2	There was a long queue at the airport.	A long queue was at the airport.
5.3	He is interested in learning Spanish.	He is interested in Spanish.
5.4	He is talking on the phone at the moment.	He is on the phone right now.
5.5	The food smells absolutely delicious.	The food is really delicious.
5.6	She made me smile with her joke.	Her joke made me smile.

In general, the results indicated that with a clear prompt, ChatGPT can generate efficient questions that focus on vocabulary and grammar (Nguyen, 2023; Zhou & Bhat, 2021).

The analysis of 120 AI-generated multiple-choice questions (MCQs) in this study revealed distinct trends in the distribution of syntactic transformations among the five ReParaphrased categories. In particular, the AI model produces fewer items in categories that require more sophisticated contextual understanding (Ell and CC) while favoring transformation types with more obvious syntactic patterns (SNC and NS).

According to the study's findings, Subordination & Nesting Changes (SNC) and Diathesis Alternation (DA) were the most common syntactic patterns, consistent with those reported by Wahle et al. (2023) and Zhou et al. (2025). In contrast, while Negation Switching (NS) was found to be the second most frequent pattern in the current research, other studies reported it as the least common.

Therefore, to effectively use AI for question generation, lecturers should provide clear prompts specifying the number and types of questions in each category, ensuring alignment with lesson goals and learning objectives.

Research question 2: What are lecturers' insights on ChatGPT's assistance in MCQ design regarding syntactic transformations?

Insights from Lecturers

The lecturers agreed that ChatGPT was effective at generating well-structured multiple-choice questions and options from the given prompts. The outputs also reflected a range of syntactic transformation patterns from the ReParaphrased framework, though the distribution across categories was uneven.

> "*I'm satisfied with the questions ChatGPT generated, as they covered all the grammar and vocabulary points in each unit. Although there were a few mistakes or awkward sentences, the questions and options were still reliable and usable.*" (L1)

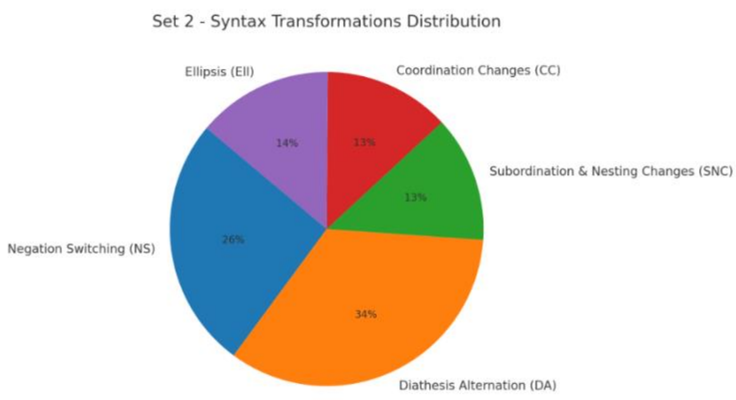
> "*Creating a large set of high-quality, non-repetitive multiple-choice questions is incredibly time-consuming while I'm trying to cover a range of grammar points and prevent students from simply memorizing patterns. An AI can take the rule for something like Diathesis Alternation and instantly generate a dozen or more valid questions. This saves me hours.*" (L2)

> *"Instead of having a single practice set for a topic, I can quickly generate multiple versions. This means I can give different practice tests to different students or even generate a unique set for each student who wants extra practice. This prevents the issue of students just memorizing the answers and truly forces them to engage with the underlying grammar rules.*" (L3)

Based on the results, the lecturers observed that ChatGPT's output was inconsistent, as it could not make decisions and its results varied over time. As a result, the second attempt yielded different findings, with Diathesis Alternation (DA) emerging as the most frequent category, rather than Subordination & Nesting Changes (SNC). However, Ellipsis (Ell) and Coordination Changes (CC) continued to be the least represented. The reason for this low preference is that the two types require advanced contextual awareness and carry a higher risk of meaning loss.

Figure 2.

Syntactic Transformation Distribution (2nd attempt)



Regarding the options, the lecturers noted that some required substantial revision before use. For example, those in the Ellipsis and Diathesis Alternation categories often contained ungrammatical or unclear sentences.

> "*I don't think all of the options are ready to use yet. Particularly, some options in the Ellipsis and Diathesis Alternation categories often come out awkward or unclear, so they need significant editing before use.*" (L1)

> "*When it comes to Ellipsis, that's where I'd have the most doubts about an AI. This type of transformation is all about taking out words that are already understood from the context, and that requires a really deep grasp of the sentence's meaning. A simple example is easy enough for an AI to handle, but for more complex sentences with intricate word gaps or tricky pronouns, I just wouldn't trust it. I'd definitely need to double-check those."* (L2)

> "*While ChatGPT is quite good at Diathesis Alternation, it often defaults to that simple active-to-passive switch. The problem is, it can churn out so many of those that some of the sentences, while grammatically correct, just don't sound like something a person would actually say. It lacks that human intuition for what sounds natural and what just sounds... a little clunky.*" (L3)

The findings from lecturers' reflections aligned with those of Nguyen (2023) and Settles et al. (2020), indicating that AI can generate accurate and efficient questions. Additionally, it is highlighted that the ChatGPT application for designing MCQs for examinations helps save time and reduce workload, with results similar to those reported in other research (Le, 2024; Nguyen, 2023). Nevertheless, Zhang and Li (2021) also emphasized that human review is essential, as results often require adjustment before they are fully reliable.

Overall, the results align with theoretical predictions that emphasize the complementary relationship between human expertise and AI capability in educational settings (Gignac & Szodorai, 2024). However, AI functions most effectively as an assistant that supports, rather than replaces, the instructor's role in designing cognitively meaningful learning experiences, since thorough supervision and deep revision of the AI-generated content are necessary before classroom use. This process involves evaluating the accuracy, appropriateness, and cognitive level of each question, and revising ambiguous or misleading items to ensure that AI-generated materials maintain quality and effectively support student learning.

Conclusion

The study explored ChatGPT's performance in generating syntax-based multiple-choice questions, revealing both its strengths and limitations. GPT-5 showed a clear preference for Subordination and Nesting Changes (SNC) and Negation Switching (NS), whereas it used Ellipsis (Ell) and Coordination Changes (CC) less often. Although the model produced questions efficiently within the theoretical framework, recurring issues such as semantic distortions and awkward double negatives or passives were observed in specific syntactic patterns. Thus, this finding highlights the importance of human intervention in proofreading and adjusting the generated questions to ensure their precision and appropriateness for use in language assessment.

While the findings of this research can be useful for understanding GPT-5's preferences when generating syntax-based multiple-choice questions, several limitations need to be addressed.

The first limitation to note is that the research has not taken into account the likelihood that GPT-5 may generate different corpora with different statistical counts for the same material, as this data-generation process is limited to two iterations per unit. Therefore, it limits the generalizability of the dominant type of syntactic change preferred by this platform. Further research is needed to replicate the procedure multiple times with the same materials to generate a corpus large enough for generalization.

Secondly, this research may have overlooked a link between the question's focus and the preferred type of syntactic transformation. As the contents of the dataset focus on grammar and vocabulary, there might be a preferred pattern of syntactic changes for grammar-focused questions and vocabulary-focused questions, as the two most dominant types of transformation, Subordination and Nesting Changes (SNC) and Negation Switching (NS), can be closely linked to a certain kind of focus. This further investigation will provide deeper insights into the transformation mechanism used by GPT-5, which can help improve prompts to reduce the likelihood of predictable or repeated patterns in certain kinds of questions.

Another limitation is the narrow scope of the research, which included only six units of Empower B1, making it insufficient to draw generalizations about the dominant types of syntactic transformation and the link between the question focus and the preferred type when the transformation is performed. The limited grammar and vocabulary may influence the transformation patterns employed by GPT-5, as some content is more likely to be transformed using one technique than others. It is also worth noting that more problems can be found in the testing items with a broader scope of grammar and vocabulary. Consequently, an additional data corpus with a broader vocabulary and grammar is needed to confirm that these results are comparable to those in other contexts.

Finally, this study employed the REParaphrased framework, focusing strictly on syntactic transformation, which may limit the diversity of transformation techniques. Specifically, because LLMs were restricted to certain kinds of structural transformations, they might have struggled to ensure that the correct answers in the test items remained semantically consistent with the original sentence while also sounding natural. This issue may not occur if LLM is allowed to be more selective in the techniques used in the generation process. During the data generation process, the prompt did not clearly specify the criteria for good paraphrasing, which might have been applied when creating the test items, so it could also adversely affect the quality of the answers generated by GPT-5. More research can be conducted by employing a wider range of transformation techniques under different frameworks, with a clearer focus on the criterion of good paraphrasing in prompt engineering, to increase awareness of the LLM's strengths and weaknesses when generating this type of question.

While ChatGPT is proven to be proficient in generating certain types of questions, alleviating their workload while promoting effective assessment, it is necessary for teachers to be aware of the patterns in terms of syntactic change preferred by the platform to refine the prompts, and the awkwardness in some certain structural transformations to make proper adjustments, aiming for a more diverse and precise generation of questions.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press, 1-305.

Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, 16-23. Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073448

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. https://doi.org/10.48550/arXiv.1803.11175

Chen, M. H., Huang, S. T., Chang, J. S., & Liou, H. C. (2015). Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning, 28*(1), 22–40. http://dx.doi.org/10.1080/09588221.2013.783873

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (2006). *Language and Mind (3rd ed.)*. New York: Cambridge Press.

Dhawaleswar Rao, C. H., & Saha, S. K. (2020). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies, 13*(1), 14–25. https://doi.org/10.1109/TLT.2018.2889100

Fries, C. C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press, 1-153.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 758–764. Association for Computational Linguistics.

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence, 104*, 101832. https://doi.org/10.1016/j.intell.2024.101832

Hamilton, C. (2025). Keyword transformations: On the border of stylistics and language testing. *Études de stylistique anglaise, 20,* 1-18. https://doi.org/10.4000/14a6j

Heaton, J.B. (1979). *Writing English Language Tests: A Practical Guide for Teachers of English*. 5th Edition, Longman, London, 138.

Hirvela, A., & Du, Q. (2013). Why am I paraphrasing? Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12*(2), 87–98. https://doi.org/10.1016/j.jeap.2012.11.005

Hosking, T., & Lapata, M. (2021). Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 1405–1418, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.112

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732980

Isley, C., Gilbert, J., Kassos, E., Kocher, M., Nie, A., Brunskill, Domingue, B., Hofman, J., Legewie, J., Svoronos, T., Tuminelli, C., & Goel, S. (2025). Assessing the Quality of AI-Generated Exams: A Large-Scale Field Study [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.08314

Kim, N., Carlson, K., Dickey, M., & Yoshida, M. (2020). Processing gapping: Parallelism and grammatical constraints. *Quarterly Journal of Experimental Psychology, 73*(5), 781-798. https://doi.org/10.1177/1747021820903461

Kovatchev, V., Martí, M. A., & Salamó, M. (2018). ETPC: A paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1384–1392.

Lado, R. (1964). *Language teaching: A scientific approach*. New York: McGraw-Hill.

Lakoff, G. (1971). On generative semantics. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*, 232–296. Cambridge: Cambridge University Press.

Lakoff, G., & Ross, J. R. (1976). Is deep structure necessary? In J. D. McCawley (Ed.), *Notes from the linguistic underground*, 159–164. Brill. https://doi.org/10.1163/9789004368859_011

Le, T. T. H. (2024). Evaluating HUFLIT Lecturers' Perspectives on ChatGPT's Capabilities in Designing English Testing and Assessment. In *Proceedings of the AsiaCALL International Conference, 6*, 157-181. https://doi.org/10.54855/paic.24612

Maas, A., Yamada, K., Nagahama, T., Kawada, T., & Horita, T. (2024). Question Generation for English Reading Comprehension Exercises using Transformers. *IIAI Letters on Informatics and Interdisciplinary Research*, *5,* 1-12. https://doi.org/10.52731/liir.v005.183

McCawley, J. D. (1968). Lexical insertion in a transformational grammar without deep structure. In *Proceedings from the 4th Annual Meeting of the Chicago Linguistic Society, 4* (1), 71–80. Chicago Linguistic Society.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence, 12*(1), 1-32. https://doi.org/10.1007/s13748-023-00295-9

Na, C. D., & Mai, N. X. N. C (2017). Paraphrasing in academic: A case study of Vietnamese learners of English. *Language Education in Asia, 8*(1), 9–24. http://dx.doi.org/10.5746/LEiA/17/V8/I1/A02/Na_Mai

Nation, I. S. P. (2009). *Teaching ESL/EFL reading and writing*. New York: Routledge.

Nguyen, T. P. T. (2023). The Application of ChatGPT in Language Test Design – The What and How. In *Proceedings of the AsiaCALL International Conference, 4*, 104-115. https://doi.org/10.54855/paic.2348

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Paribakht, T. S. (2004). The role of grammar in second language lexical processing. *RELC Journal, 35*(2), 149-160. https://doi.org/10.1177/003368820403500204

Poppels, T. (2020). *Towards a referential theory of ellipsis*. University of California, San Diego, 1-242. https://escholarship.org/uc/item/2830w1xn

Postal, P. M. (1974). *On raising: One rule of English grammar and its theoretical implications*. Cambridge, MA: MIT Press.

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end Generation of Multiple-choice Questions using Text-to-text Transfer Transformer Models. *Expert Systems with Applications,* 118258. https://doi.org/10.1016/j.eswa.2022.118258

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine Learning - Driven Language Assessment. Transactions of the Association for Computational Linguistics, 8, 247-263. https://doi.org/10.1162/tacl_a_00310

The Case HQ. (2025, April 7). *Powerful guide to writing exam questions using Gen AI effectively*. The Case HQ. https://thecasehq.com/powerful-guide-to-writing-exam-questions-using-gen-ai-effectively/

Thompson, D., Ling, S. P., Myachykov, A., Ferreira, F., & Scheepers, C. (2013). Patient-related constraints on get- and be-passive uses in English: evidence from paraphrasing. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00848

Tran, T. T. T., & Nguyen, H. B. (2022). The Effects of Paraphrasing on EFL Students' Academic Writing. *Journal of Language and Linguistic Studies. 18*(1), 976-987.

Vahtola, T., Creutz, M. & Tiedemann, J. (2022). It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark. In *Proceedings of the 5th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP,* 249–262. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.blackboxnlp-1.20

Vila, M., Martí, M. A., & Rodríguez, H. (2014). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics, 4*(3), 205–218. https://doi.org/10.4236/ojml.2014.41016

Wahle, J. P., Gipp, B., & Ruas, T. (2023). Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12148-12164. https://doi.org/10.18653/v1/2023.emnlp-main.746

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Wieting, J., & Gimpel, K. (2017). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistic, 1,* 451-462. https://doi.org/10.18653/v1/P18-1042

Wieting, J., Bansal, M., Gimpel, K., Livescu, K., & Roth, D. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics, 3*, 345–358. https://doi.org/10.1162/tacl_a_00143

Zhang, M., & Li, J. (2021). A Commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research, 1*(6), 831-833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT [Preprint]. ArXiv (Cornell University). https://doi.org/10.48550/arXiv.1904.09675

Zhou, C., Qiu, C., Liang, L., & Acuna, D. E. (2025). Paraphrase Identification with Deep Learning: A Review of Datasets and Methods. *IEEE Access*, *13*, 65797-65822. https://doi.org/10.1109/access.2025.3556899

Zhou, J., & Bhat, S. (2021). Paraphrase Generation: A Survey of the State of the Art. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5075–5086. https://doi.org/10.18653/v1/2021.emnlp-main.414

Zhou, Z., Sperber, M., & Waibel, A. (2019). Paraphrases as Foreign Languages in Multilingual Neural Machine Translation [Preprint]. ArXiv (Cornell University). https://doi.org/10.18653/v1/p19-2015

Biodata

Vu Thi Kim Chi is currently a full-time lecturer at the Faculty of Foreign Languages - Saigon University. She received a Master's degree in TESOL from Victoria University, Australia. Her teaching practice involves language skills courses attended by non-English major students. She has sought to apply technological innovation in the classroom to improve teaching and learning outcomes.

Nguyen Trinh To Anh has been an EFL instructor at the Faculty of Foreign Languages, Saigon University, Vietnam, since 2015. She holds a Master’s degree in English Language Studies from Hoa Sen University, Vietnam; a Master of Commerce specializing in Tourism and Hospitality Management from Macquarie University, Australia; and a Master of Professional Accounting from La Trobe University, Australia. Her teaching responsibilities include English for General Purposes and English for Specific Purposes, particularly in the fields of Accounting – Auditing, Finance – Banking, and Business Management. Her research interests focus on ESP teaching methodology and blended learning approaches.

Vo Dao Vuong Co is currently a full-time lecturer at the Faculty of Foreign Languages – Saigon University. She received a Master’s degree in Applied Linguistics from Curtin University, Australia. Her teaching experience includes test-prep courses, general English courses for non-English-major students, and linguistics courses for English-major students. She has aimed to improve the learning experience and teaching outcome by creating a classroom where learning meets innovation.