UDC 37.091.2:004.8:811.111

**Olena Zhyhadlo**
PhD in Linguistics, Associate Professor, Associate Professor at Foreign Languages Department, Educational and Scientific Institute of Law
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID ID 0000-0002-1605-7242
*olena.zhygadlo@gmail.com*

**Inna Zaiarna**
PhD in Education, Associate Professor, Associate Professor at Foreign Languages Department, Educational and Scientific Institute of Law
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine;
Leading Researcher at Comparative Studies Department for Information and Education Innovations
Institute for Digitalisation of Education of the National Academy of Educational Sciences of Ukraine, Kyiv, Ukraine
ORCID ID 0000-0002-9464-096X
*izaiarna@gmail.com*

# ARTIFICIAL INTELLIGENCE-DRIVEN TESTING IN EFL/ESP CLASSROOMS

**Abstract.** This article explores the application of Artificial Intelligence (AI)-driven tools, particularly ChatGPT, for creating vocabulary test tasks in EFL/ESP classrooms. The research aims to evaluate the quality of vocabulary test tasks generated by ChatGPT by applying established criteria, including relevance, reliability, interactiveness, practicality, and impact. It investigates how ChatGPT-generated tasks meet these criteria and provides practical recommendations for educators to optimize the quality of AI-generated assessments. The authors indicate that criteria such as relevance, practicality, interactivity, and impact can be fully satisfied in ChatGPT-generated tests. However, the research identifies challenges with the reliability of AI-generated test tasks, primarily due to ambiguities in response choices.

The article emphasizes the pivotal role of human intervention in guiding and refining AI-generated outputs. Detailed and context-specific prompts crafted by educators are critical to maximizing the potential of ChatGPT while mitigating its limitations. To support EFL/ESP teachers, the study offers detailed recommendations for enhancing ChatGPT-generated test tasks, such as developing precise prompts, setting clear contexts, assigning specific roles to ChatGPT, and iteratively refining outputs. These strategies improve the reliability and effectiveness of AI-generated assessments and align them with pedagogical standards. The authors emphasise the importance of integrating human oversight with AI tools to maintain the validity and usefulness of language tests. This research contributes to the broader discourse on integrating AI in education by demonstrating how educators can leverage ChatGPT for test design while addressing its limitations. Future directions include evaluating the effectiveness of other types of AI-generated test tasks, exploring AI’s role in automated assessment and feedback, and examining the long-term impact of AI-driven assessments on teaching methodologies and students’ vocabulary acquisition in ESP contexts.

**Keywords:** Artificial Intelligence; ChatGPT; Vocabulary Test Design; EFL/ESP Classrooms; AI-Generated Assessments; Language Testing Criteria.

## 1. INTRODUCTION

**Statement of the problem.** Artificial Intelligence (AI)-driven tools, such as ChatGPT, are increasingly being used in educational settings, particularly for automating test creation. This trend reflects the broader integration of AI technologies in educational practices, where the potential to streamline tasks such as content generation and assessment design holds considerable appeal. However, while these AI tools are capable of performing certain language-related tasks with remarkable speed, it is advisable to utilize their output with caution due to bias or false information it may potentially contain.

Recent literature highlights that automated AI-generated content often lacks the nuanced understanding of linguistic or contextual intricacies that human educators bring to task design [1]. Furthermore, concerns have been raised about the ethical implications and practical limitations of relying heavily on AI tools in high-stakes educational contexts, including the potential for perpetuating biases and inaccuracies [2]. For example, a survey on the application of ChatGPT for assessment in English language teaching, which involved EFL and ESP instructors from higher education institutions across different countries, revealed that the majority of participants who attempted to use this AI tool in their teaching practice expressed concerns regarding the accuracy of ChatGPT-generated tasks and reported issues with the presence of inaccurate and biased information in the generated content [3].

Given the increasing reliance on AI tools in education, it becomes imperative to establish frameworks for ensuring the quality of their outputs, particularly in the domain of language assessment. Misaligned or poorly constructed test tasks can compromise the validity and reliability of assessments, undermining their pedagogical effectiveness. Therefore, ensuring the quality of ChatGPT-generated tasks is essential for maintaining assessment reliability and aligning them with pedagogical standards.

**Analysis of recent research and publications.** The integration of AI into foreign language education has been the subject of numerous studies, reflecting its growing relevance in educational practice. Researchers have explored various aspects of AI applications, including automated test generation, adaptive learning systems, and interactive tools that facilitate vocabulary acquisition and language practice. To fulfill the objectives of this study, our analysis focuses only on recent research examining the use of AI tools, such as ChatGPT, in EFL and ESP contexts.

Recent studies underscore the potential of AI to enhance foreign language learning and teaching. For instance, the systematic review by O. Nalyvaiko et al. [4] highlights the transformative role of AI in learning management systems, particularly for developing language skills such as speaking, writing, and listening. The study emphasizes the personalized learning experiences AI facilitates, which adapt to individual learner needs and provide engaging, interactive environments.

In another publication, A. Kyrpa et al. [5] investigate the application of AI tools in social sciences and humanities, including language education. This study examines educators' competencies in utilizing AI technologies like ChatGPT for lesson planning and content creation, offering practical examples and recommendations for effective integration. The authors also discuss challenges such as the risk of inaccuracies in AI-generated content and the importance of aligning these tools with pedagogical standards.

Further, a study by Schmidt and Strasser [6] explores the potential of AI in computer-assisted language learning, emphasizing intelligent tutoring systems and adaptive learning technologies that cater to individual learners. The findings suggest that while AI offers significant advantages in personalized education, human oversight must complement its use to address contextual and cultural nuances.

Karataş et al. [7] also conducted a qualitative case study investigating ChatGPT's impact on foreign language learners. Their research provides detailed insights into the tool's practical application, highlighting both its advantages in offering contextualized and interactive learning opportunities and its limitations in delivering consistent accuracy.

A systematic review published by the British Council [8] studies the role of AI in English language teaching, analyzing its potential to enhance assessment, task design, and learner engagement. However, the study also raises concerns about ethical issues and the possibility of AI "hallucinations" resulting in misleading outputs.

Another notable contribution is the work of A. Mugableh [9], which evaluates the impact of AI-driven chatbots on vocabulary acquisition in EFL classrooms. The study demonstrates how tools like ChatGPT can improve vocabulary retention by providing contextualized and interactive learning experiences. However, it also cautions against over-reliance on AI, emphasizing the need for teacher mediation to ensure the accuracy and relevance of learning materials.

These studies collectively highlight the transformative potential of AI in language education while drawing attention to its limitations. The findings provide a foundation for developing criteria and recommendations for effectively incorporating AI tools like ChatGPT into EFL and ESP classrooms. This analysis underscores the importance of human oversight in leveraging AI's capabilities to align with pedagogical objectives and enhance learner outcomes.

The **purpose** of this research is to explore the capabilities of AI to create test tasks for assessing specialized vocabulary in the English language classroom and to develop a method for evaluating the quality of vocabulary test tasks generated by ChatGPT. To achieve this goal, the following objectives have been set – to establish the criteria for evaluating test tasks generated by ChatGPT based on input from a human test designer and to develop recommendations for EFL/ESP teachers involved in designing vocabulary tests on how to enhance the quality of the chatbot's output. By addressing these objectives, this study aims to contribute to the growing discourse on the integration of AI in education by offering actionable insights that enable educators to maximize the benefits of these tools while mitigating their limitations. Ultimately, by developing criteria and recommendations, the study will empower educators to enhance the effectiveness of AI in language testing, benefiting both teachers and learners.

## 2. RESEARCH METHODS

The authors employed various theoretical research methods, including critical analysis of existing scholarly works on AI's application in teaching and learning, as well as abstraction and specification, comparison, and analogy methods – to draw analogies between AI-generated tests and human-created tests. The empirical method of analysing the authors' own endeavours of ChatGPT-assisted test tasks development was used to compile recommendations on improving the quality of the AI-generated output for English language instructors.

Based on the premise that generative AI is not merely a tool but an integral component of test creation, the study applied principles of language assessment traditionally used for human-developed vocabulary tests to evaluate the effectiveness of ChatGPT-generated vocabulary tests.

## 3. FINDINGS

Describing test taking as a tedious and nerve-wracking experience can hardly be contested either by those who have taken any kind of achievement test or by those who develop them. Regardless of their format, tests should aim to measure test-takers' knowledge, ability and performance in a way that the results could provide testees with constructive feedback, motivate further learning, and enable testers to tailor instruction to students' needs and abilities. This approach makes the assessment process beneficial for both parties. In addition, the issue of time- and cost-effectiveness of creating, administering, and taking a test must be considered, as test developers and classroom teachers do not have unlimited time at their disposal. The effort invested in designing a classroom achievement test should be reasonable and not excessive.

Generative AI, particularly through large language models, can produce text that closely resembles human writing. This capability allows us to regard it not merely as a tool but as an integral part of the writing process – essentially a co-creator [10]. ChatGPT, a natural language processing tool, excels in generating highly targeted content based on prompts and continuously refining its tutoring capabilities, which makes it an inviting assistant for instructors. One of its applications is in language teaching and learning.

However, while utilizing the potential of ChatGPT to create tests that assess knowledge, skills, and performance of specific students within a particular language domain in EFL or ESP classroom appears highly appealing, the resulting products must be approached with caution, just like human-developed tests and assignments. This suggests that a set of rigorous criteria should be established to assess the quality of the tests generated by AI.

In this research, we explore the potential of ChatGPT to generate assignments that can be used to evaluate vocabulary knowledge in the English language classroom. The criteria applied in this study to assess the quality of a language test developed by ChatGPT-4 following the instructions of a professional language teacher are based on the principles of language assessment [11, p. 33] and the characteristics of "usefulness" of a language test elaborated by L. Bachman and A. Palmer in [12, p. 17] which are traditionally used to evaluate human-developed language tests.

L. Bachman and A. Palmer defined usefulness as "a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test" [12, p. 18]. In this research, usefulness of a vocabulary test developed with the assistance of ChatGPT is construed as partial or complete fulfilment by the test of all the requirements or criteria set for a human-developed vocabulary test. The diagram below illustrates a set of specific criteria applied to a vocabulary test as part of formative assessment that needs to be met for the test to be considered useful (Figure 1).

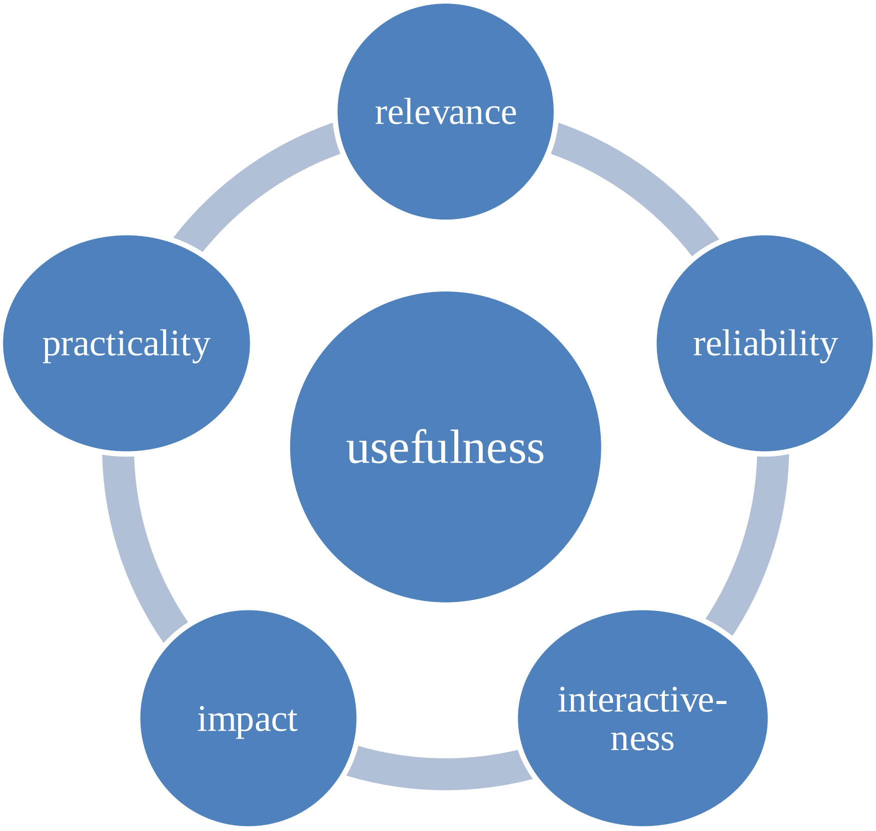


*Figure 1. Criteria of Vocabulary Test Usefulness*

Complete or partial fulfilment of all the above criteria is essential for a vocabulary test to be deemed useful. The order in which the criteria are arranged in the diagram and described does not imply any priority.

Relevance or authenticity in Bachman and Palmer's terms reveals how the test task correlates with the target language use. According to Brown [11, p. 28], relevance or authenticity is manifested in the language material used in the test, which should be as natural as possible, in contextualized items, in the topics which are relevant for the learner, in certain thematic organization of the test items, in tasks imitating real-life tasks. For example, test items designed to be used in the lessons of Legal English should be based on the texts of current legislation, contain up-to-date legal terminology, and contain provisions that go in line with contemporary socio-legal theories.

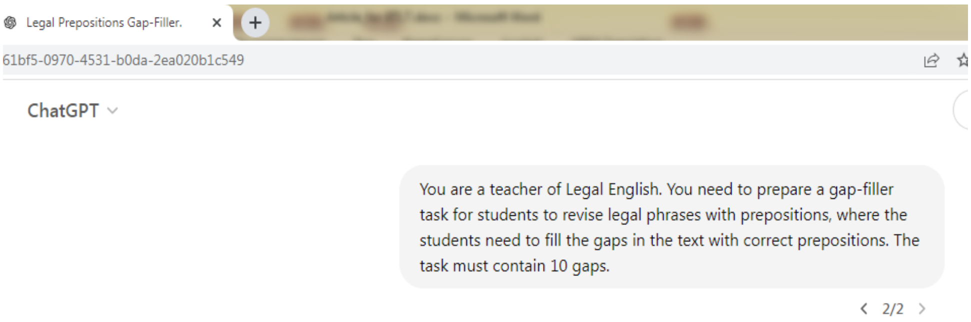
Reliability of a language test is connected with its consistency and dependability in terms of results demonstrated by the same student or different students on different occasions and when assessed by different raters [11, p. 21]. This is closely related to the quality of scoring, human factor, potential subjectivity, bias, and test administration conditions. In addition to these factors, the design of the test – including its length, the description of test items, and the ambiguity of answers provided in a multiple-choice format – also significantly impacts its reliability.

The practicality of a vocabulary test is associated with the time and effort invested by human test creators in its design and the convenience and efficiency of its implementation, including the time spent by test takers on its completion, and can be evaluated in terms of cost-effectiveness and time-efficiency. A cost-effective test should not impose an excessive financial burden on test creators and is easy to administer. The test should be time-efficient as regards the time spent on its creation, testing, proofreading and subsequent scoring/evaluation. Besides, test creators should ensure the possibility of completing the test reasonably within the allotted time frame. A vocabulary test created with the assistance of ChatGPT-3.5 can be considered practical if a human test maker spends a reasonable amount of time not only on formulating the prompt for AI to produce the expected result but also on evaluating the usefulness of the created test. Though the term "reasonable time" may seem rather vague, it should be interpreted as time that is necessary to do the task appropriately and conveniently.

As far as interactiveness is concerned, Bachman and Palmer define it "as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" [12, p. 25]. In language testing, the following characteristics can be considered relevant: language knowledge, metacognitive strategies, and topical knowledge. Thus, unless a test task requires the test takers to use their knowledge of the target language domain and the knowledge of the relevant topic and apply metacognitive strategies such as goal-setting, planning, monitoring mistakes, etc., it cannot be considered interactive. For instance, in the context of Legal English lessons, tasks such as substituting formal words with their plain English equivalents in a letter of advice to a client or filling in the gaps in an excerpt from a contract with technical terms would be both relevant and interactive as these activities closely resemble the tasks lawyers encounter in their day-to-day work and allow the teacher to draw inferences about the test takers' language proficiency and subject matter knowledge.

Finally, impact or washback [11, p. 29] is the criterion that represents "the effect of testing on teaching and learning" [13, p. 1], which can be either beneficial or harmful. In our opinion, a positive impact of testing on students is achieved through constructive and meaningful feedback provided to test takers as part of the assessment process, indicating areas of strength and areas that require improvement. This feedback, presented as a score, influences test takers' perception of their language abilities, potentially enhancing their external motivation to engage more deeply with a topic. Another issue that may maximize the positive impact on students is the quality of the materials used as the basis for test tasks. If task descriptions are well-written and not ambiguous, the test items provide new information which is interesting for test takers from the professional perspective, and the test meets their expectations in terms of task format and language proficiency level, the experience of taking it may be quite instructive, motivating the test takers to learn not only from feedback but while actually taking the test. At the same time, teachers can also benefit from washback, as they receive valuable information about students' performance, which enables them to adjust their teaching strategies to the objective needs of their students.

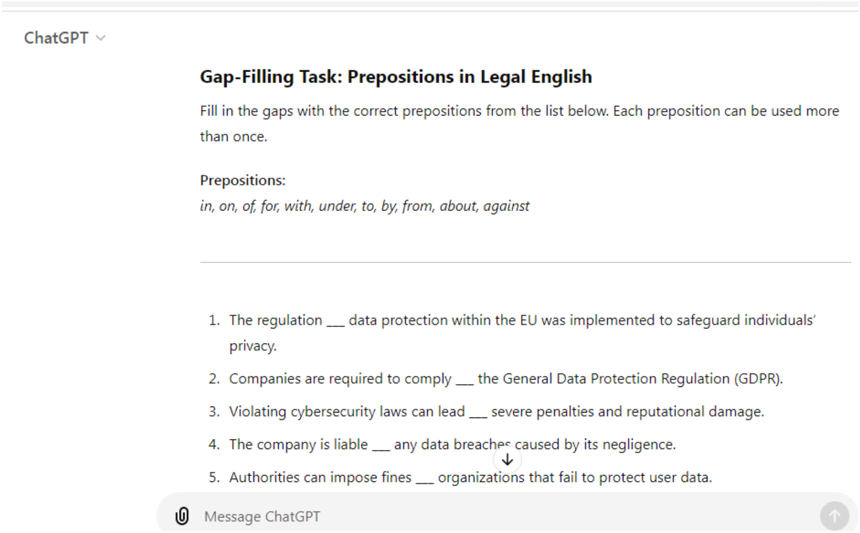
Among the assignments used to assess target vocabulary in ESP lessons, we have selected gap-filling tasks for analysis, as they have proved effective for practicing specific language points and are widely used in testing. These tasks are advantageous because they allow for targeted assessment of contextual vocabulary knowledge, are adaptable to various proficiency levels and professional fields, encourage active recall, helping students reinforce word meaning and usage in a meaningful context. Additionally, it is relatively easy to create gap-fill tasks, which consist of separate sentences or excerpts containing gaps to evaluate students' vocabulary production, with the help of ChatGPT. Below, there is the fragment of a log with the prompt used by a human task creator for ChatGPT to design the gap-fill task (Figure 2), specifying the context, target vocabulary, type of the activity, and number of sentences.


Legal Prepositions Gap-Filler.
61bf5-0970-4531-b0da-2ea020b1c549
ChatGPT
2/2

You are a teacher of Legal English. You need to prepare a gap-filler task for students to revise legal phrases with prepositions, where the students need to fill the gaps in the text with correct prepositions. The task must contain 10 gaps.

*Figure 2. The prompt for the gap filling task*

Figure 3 illustrates the task for the gap fill exercise generated by ChatGPT in response to the prompt, which has been used by a human test developer to create a selected-response task with the prepositions arranged in the box for students to choose from while filling in the blanks. In the output, ChatGPT provides the task description, specifying that each key phrase can be used more than once. As a response to a prompt, ChatGPT offers the key to the task.



**Gap-Filling Task: Prepositions in Legal English**

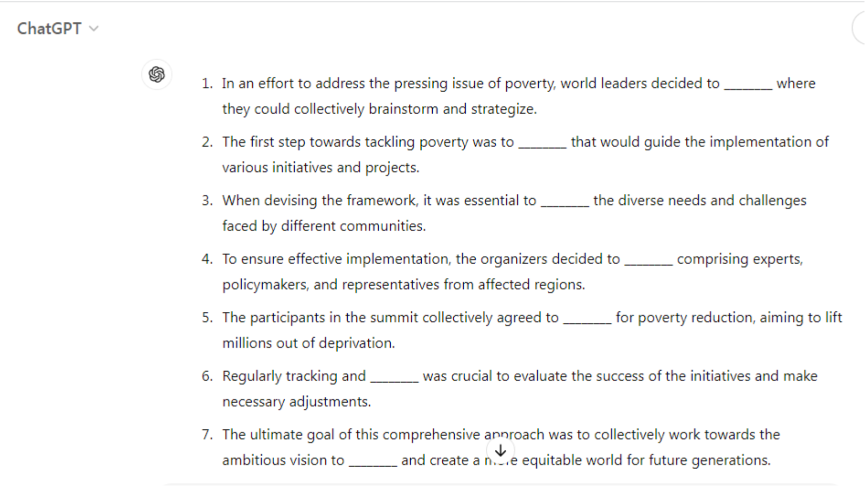
Fill in the gaps with the correct prepositions from the list below. Each preposition can be used more than once.

**Prepositions:**
*in, on, of, for, with, under, to, by, from, about, against*

1. The regulation ___ data protection within the EU was implemented to safeguard individuals' privacy.
2. Companies are required to comply ___ the General Data Protection Regulation (GDPR).
3. Violating cybersecurity laws can lead ___ severe penalties and reputational damage.
4. The company is liable ___ any data breaches caused by its negligence.
5. Authorities can impose fines ___ organizations that fail to protect user data.

*Figure 3. Task description generated by ChatGPT*

Another example of a vocabulary task generated by ChatGPT in response to the prompt to create a gap-filling exercise containing the following selected key phrases: "host a summit, establish framework, take into consideration, set up groups, set targets, measure progress, end poverty", where each key phrase can only be used once, is presented in Figure 4. This task aims to check the knowledge of International Law vocabulary in a Legal English class for undergraduate law students.



1. In an effort to address the pressing issue of poverty, world leaders decided to ______ where they could collectively brainstorm and strategize.
2. The first step towards tackling poverty was to ______ that would guide the implementation of various initiatives and projects.
3. When devising the framework, it was essential to ______ the diverse needs and challenges faced by different communities.
4. To ensure effective implementation, the organizers decided to ______ comprising experts, policymakers, and representatives from affected regions.
5. The participants in the summit collectively agreed to ______ for poverty reduction, aiming to lift millions out of deprivation.
6. Regularly tracking and ______ was crucial to evaluate the success of the initiatives and make necessary adjustments.
7. The ultimate goal of this comprehensive approach was to collectively work towards the ambitious vision to ______ and create a more equitable world for future generations.

*Figure 4. The vocabulary task generated by ChatGPT*

To illustrate how the criteria of test usefulness can be applied to evaluate the quality of AI-designed tests, we have analysed the above vocabulary task. The results of the analysis are presented in Table 1. The first column of the rubric lists the usefulness criteria, and the top row represents the levels of criteria satisfaction for this task.

*Table 1.*

*Evaluation of usefulness of the vocabulary task (gap fill)*

| | **Fully met** | **Partially met** | **Not met** |
|---|---|---|---|
| **Relevance** | + | | |
| **Reliability** | | + | |
| **Interactiveness** | + | | |
| **Impact/washback** | + | | |
| **Practicality** | + | | |

The content of the sentences generated for the task seems relevant as it aligns with general objectives and on-going initiatives of the UN, which can be checked with the official UN sources and latest reports on the issues of sustainable development. It also correlates with the syllabus topic of International law and the materials processed by students on the role of the UN as a central platform for promoting the development and implementation of international legal norms. Thus, the criterion of relevance is fully met.

When analyzing the reliability of a separate test task, at least two things should be taken into account: task description and the issue of ambiguity as regards the choice of the response for each of the sentences, as test takers may make a different but no less valid choice than the one anticipated by the test developer (in our case AI). The task description, in this case, is fairly detailed and clear. As for ambiguity, in this task, for example, the same key phrase "set targets" seems to fit the context both in sentence 2 (where the correct answer is "establish framework" and 5, where it is the correct answer). In sentence 5, however, there is a clue in the form of a synonym, "aiming," which can be disregarded by students in the absence of any other limitation. Thus, the criterion of reliability is only partially met.

If dissatisfied with the variant provided by ChatGPT, the human test developer can select the "Try again" option in the dropdown list to be offered an improved version of the task within seconds. This will result in spending additional time on evaluating the new output, which may affect the criterion of practicality. On the other hand, the time spent by the human test developer on searching various resources on the internet for suitable topic-related phrases to transform them into a gap-fill task may significantly exceed the time spent by the same test developer on evaluating the quality of responses generated by AI. So, for creating a selected-response gap-fill task, the practicality criterion seems to be fully met from the point of view of the time and effort invested in its design by the human test developer and the resources employed.

The next criterion to be considered is interactiveness of the test task. The gap-fill task in question may be considered interactive as it allows the teacher to make inferences about the test takers' knowledge of the key vocabulary of the topic, their ability to analyze the context and utilize the right key phrases, which, among other things, involves activating prior knowledge of the UN activities within the framework of international law. Other metacognitive strategies beneficial for performing a gap-fill task are planning your approach to performing the task, predicting possible answers, monitoring your decision-making process, using context clues, and reviewing.

The last but not least characteristic of the test task is its impact on students and their learning process and the instruction. After performing the task, the students will receive some sort of feedback from the teacher, which may include their score representing the number of correct answers as well as a verbal description of their performance and recommendations on how to improve it during the analysis of the task in class. This aspect of washback depends exclusively on the teacher and students, so the teacher needs to think about how to make the feedback as meaningful and useful for their students as possible, and students need to take effort to digest it, which is essential to develop positive response to the task on their part.

There is another factor that may implicitly affect the washback on test takers either positively or negatively. This factor is related to the ambiguity of answers, which can be considered equally correct. In high-stakes tests, the presence of such ambiguity may deteriorate the result and have a harmful effect on the test taker. In lower-stakes tests, however, if they are used as formative assessment and are analysed in class, students may find it motivating and instructive if they manage to assess the options, prove their choice, and even point out the flaw to the teacher. When the reliability of the given gap-fill task was discussed earlier in the article, the ambiguity of this kind was revealed. As a result, even if the reliability criterion is not fully satisfied due to the presence of ambiguity, the impact criterion may be completely met. In case students are active participants in the assessment process, detecting and discussing ambiguities in the test task may have a beneficial impact on students' learning.

On balance, as regards the given gap-fill task, all the criteria of usefulness are either fully or partially met, which makes it possible to make a conclusion about the relative usefulness of this particular test task.

The quality of the output generated by AI tools is directly related to the effort that a human task designer invests in prompt engineering. Prompt engineering is the process of creating and refining instructions for an AI tool. To effectively guide ChatGPT to generate a test task that is finely tuned to the course requirements and lesson content and that can reliably assess the test-takers' vocabulary knowledge, the researchers have compiled recommendations for English language teachers, based on tips from Type, the AI document editor resource [14] and their own practical experience.

A prompt, i.e. a set of instructions for AI to generate a test task, should be in the form of a detailed request that specifies the following issues:

- type of a task, e.g. "please create a gap-filling exercise" or "create a true-false test";
- audience, e.g. "create a vocabulary task for law students in the master's program" or "prepare a lead-in task to introduce the topic of the rule of law to senior law students";
- purpose, e.g. "you are a Legal English teacher and you need to create a vocabulary task to assess vocabulary knowledge in Criminal law";

- scope of the output, e.g. “the task must include 10 sentences” or “create a task for the class of Academic Writing consisting of a 300-word excerpt from an academic text which must include instances of vague language”;
- context of the textual material used as the basis for the task, e.g. “create a task for the class of Academic Writing aiming to illustrate poor academic style instances and encouraging the students to improve them. The task must be based on a jurisprudence-related academic text”.

To give the context for the task, it may be advisable to assign ChatGPT a role or persona. Beginning your instructions with a phrase such as “You are a Legal English teacher” or “You are an instructor of the course in Academic Writing for PhD students in Law” may significantly enhance the accuracy and the focus of the output.

It is noteworthy that ChatGPT possesses an ability to retain the context of previous instructions within a given conversation log. This feature allows for a more seamless interaction, enabling educators to build upon previous prompts without the need for constant repetition. As a result, teachers can develop comprehensive and coherent test materials while maintaining continuity in the AI’s understanding of the task at hand. This contextual memory facilitates a smoother workflow and enhances the overall process of prompt-based test generation.

Another valuable feature of ChatGPT is its ability to generate refined versions of previously provided outputs. If educators are dissatisfied with the initial version of a vocabulary test or specific content, they can request modifications by providing more detailed instructions or clarifying their expectations. Teachers should specify the aspects they wish to adjust, such as the difficulty level, formatting, or types of vocabulary questions. By doing so, the model can effectively tailor the output to better align with pedagogical goals, offering enhanced flexibility and precision in content creation.

In addition, in the ChatGPT interface, there is a “Try again” function, usually available in the dropdown menu next to each response. This feature allows users to quickly regenerate a response without having to give specific verbal instructions. This function can be particularly useful when the initial output is not fully satisfactory, but the user prefers not to provide detailed feedback or specific instructions for further refinement.

## 4. CONCLUSIONS

The research explores the potential of AI in generating test tasks for assessment in English language lessons, particularly vocabulary tasks created by ChatGPT with the help of prompts by a language teacher. To evaluate the quality of ChatGPT-driven tests, specific criteria designed for human-developed vocabulary tests have been applied. Following Bachman and Palmer’s definition, these criteria correlate with a set of qualities referred to as “usefulness” of a language test, which includes relevance, reliability, interactiveness, practicality and impact. Whether a vocabulary test completely or partly meets the above criteria can prove its “usefulness” or, in other words, its effectiveness as a means of assessment.

The authors analyse the way each criterion of usefulness can be applied to ChatGPT-generated test tasks in order to evaluate their effectiveness in testing vocabulary knowledge. Relevance or authenticity ensures that test tasks reflect real-life language use through natural, contextualized materials, relevant topics, and tasks, such as using current legal texts and terminology for Legal English assessments. Reliability of a language test refers to the consistency of results across students, occasions, and testers and is influenced by scoring quality, human factors, test administration conditions, and test design elements such as length, item clarity, and answer ambiguity. Practicality of a vocabulary test involves cost-effectiveness, time-efficiency, and ease of implementation, with emphasis on the time spent by human test creators in formulating prompts for AI, evaluating, proofreading, and ensuring the test can be completed within a reasonable timeframe. The interactiveness of a test task requires the test takers to use their knowledge of the target language domain and apply metacognitive strategies such as goal-setting, planning, and monitoring mistakes. The quality of test materials, including clear task descriptions, professionally relevant content, and appropriate task format, can create a positive impact on the students while providing teachers with valuable washback to adjust their teaching strategies effectively.

Applying these criteria to evaluate the usefulness of gap-fill tasks showed that the criteria of relevance, practicality, interactivity, and impact can be fully satisfied in a ChatGPT-generated test. The reliability criterion, however, may not be fully satisfied due to occurrence of ambiguity in response choices. Although violation of the reliability criterion in the form of ambiguities may be considered a serious flaw causing a negative effect on test takers in high-stake tests, this issue may not impair the quality of tasks in formative assessment, as students may consider detecting and analyzing ambiguities instructive and beneficial for learning.

It has been established that the effectiveness of AI-generated output largely depends on the precision of prompts provided by a human task designer. A number of recommendations have been outlined to ensure the quality of ChatGPT-generated tests, including creating prompts in the form of detailed requests, setting the context, assigning ChatGPT a specific role, capitalizing on the capacity of ChatGPT to retain the context of previous instructions, and generating multiple refined versions of the previously provided output.

As the research primarily focused on examining the process of creating gap-fill tasks and gauging their usefulness with the help of elaborated criteria, the prospects of future investigation may include analysing the effectiveness of other types of ChatGPT-generated test tasks. Additionally, it would be advisable to explore the potential of AI to assess students' test papers based on the set criteria and provide constructive feedback. Another promising area for research lies in examining the long-term washback effects of AI-generated assessments on teaching methodologies and students' vocabulary acquisition in ESP settings. Furthermore, further studies could compare the findings of this research with those of other studies on AI-assisted test creation to identify common patterns, potential limitations, and best practices for optimizing AI-generated assessments.

## REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review," *Smart Learn. Environ.*, vol. 11, p. 28, 2024. [Online]. Available: https://doi.org/10.1186/s40561-024-00316-7 (in English)

[2] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," *AI Ethics*, vol. 2, pp. 431–440, 2022. [Online]. Available: https://doi.org/10.1007/s43681-021-00096-7 (in English)

[3] I. Zaiarna, O. Zhyhadlo, and O. Dunaievska, "ChatGPT in Foreign Language Teaching and Assessment: Exploring EFL Instructors' Experience," *ITLT*, vol. 102, no. 4, pp. 176–191, Sep. 2024. doi: 10.33407/itlt.v102i4.5716. (in English)

[4] O. Oluwafemi Ayotunde, D. I. Jamil, and N. Cavus, "The Impact of Artificial Intelligence in Foreign Language Learning Using Learning Management Systems: a Systematic Literature Review," *ITLT*, vol. 95, no. 3, pp. 215–228, Jun. 2023. doi: 10.33407/itlt.v95i3.5233. (in English)

[5] A. Kyrpa, O. Stepanenko, V. Zinchenko, T. Datsiuk, I. Karpan, and N. Tilniak, "Artificial Intelligence Tools in Teaching Social and Humanitarian Disciplines," *ITLT*, vol. 100, no. 2, pp. 162–179, Apr. 2024. doi: 10.33407/itlt.v100i2.5563. (in English)

[6] T. Schmidt and T. Strasser, "Artificial Intelligence in Foreign Language Learning and Teaching: A CALL for Intelligent Practice," *Anglistik: International Journal of English Studies*, vol. 33, no. 1, pp. 165–184, Spring 2022. doi: 10.33675/ANGL/2022/1/14. (in English)

[7] F. Karataş, F. Y. Abedi, F. O. Gunyel, et al., "Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners," *Educ Inf Technol*, vol. 29, pp. 19343–19366, 2024. [Online]. Available: https://doi.org/10.1007/s10639-024-12574-6 (in English)

[8] H. Crompton, A. Edmett, and N. Ichaporia, "Artificial intelligence and English language teaching: A systematic literature review," *British Council*, 2023. [Online]. Available: https://www.britishcouncil.org/sites/default/files/ai_in_english_language_teaching_systematic_review.pdf?utm_source=chatgpt.com (in English)

[9] A. I. Mugableh, "The Impact of ChatGPT on the Development of Vocabulary Knowledge of Saudi EFL Students," *Arab World English Journal (AWEJ)*, Special Issue on ChatGPT, pp. 265–281, Apr. 2024. doi: https://dx.doi.org/10.24093/awej/ChatGPT.18. (in English)

[10] K. K. Davis, "A New Parlor is Open: Legal Writing Faculty Must Develop Scholarship on Generative AI and Legal Writing," *Stetson Law Review Forum*, vol. 7, no. 1, pp. 1, 2024. [Online]. Available: https://www2.stetson.edu/law-review/article/a-new-parlor-is-open-legal-writing-faculty-must-develop-schola. (in English)

[11] H. Brown, *Language Assessment: Principles and Classroom Practices*, White Plains, NY: Longman, 2004, p. 33. (in English)

[12] L. Bachman and A. Palmer, *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford: Oxford University Press, 1996. (in English)

[13] A. Hughes, *Testing for Language Teachers*, 2nd ed., Cambridge: Cambridge University Press, 2003.

[14] *How to write AI prompts that get results*. [Online]. Available: https://blog.type.ai/post/how-to-write-ai-prompts-that-get-results. [Accessed: Dec. 30, 2024]. (in English)

*Text of the article was accepted by Editorial Team 30.12.2024*

# ШТУЧНИЙ ІНТЕЛЕКТ У ТЕСТУВАННІ З АНГЛІЙСЬКОЇ ЯК ІНОЗЕМНОЇ ТА ДЛЯ СПЕЦІАЛЬНИХ ЦІЛЕЙ

**Олена Жигадло**
кандидат філологічних наук, доцент, доцент кафедри іноземних мов Навчально-наукового інституту права
Київський національний університет імені Тараса Шевченка, м. Київ, Україна
ORCID ID 0000-0002-1605-7242
*olena.zhygadlo@gmail.com*

**Інна Заярна**
кандидат педагогічних наук, доцент, доцент кафедри іноземних мов Навчально-наукового інституту права
Київський національний університет імені Тараса Шевченка, м. Київ, Україна;
провідний науковий співробітник відділу компаративістики інформаційно-освітніх інновацій
Інститут цифровізації освіти НАПН України, м. Київ, Україна
ORCID ID 0000-0002-9464-096X
*izaiarna@gmail.com*

**Анотація.** Ця стаття присвячена особливостям застосування технологій на основі штучного інтелекту (ШІ), зокрема ChatGPT, для створення завдань, націлених на перевірку рівня сформованості лексичних навичок студентів на заняттях з англійської мови як іноземної та/або для спеціальних цілей. Дослідження спрямоване на оцінювання якості завдань, згенерованих ChatGPT, за допомогою визначених критеріїв, таких як відповідність, надійність, інтерактивність, практичність та вплив. У статті проаналізовано, наскільки тестові завдання, створені ChatGPT, відповідають цим критеріям, та надано практичні рекомендації для викладачів щодо покращення якості тестових завдань, згенерованих ШІ. Автори зазначають, що такі критерії, як відповідність, практичність, інтерактивність та вплив, можуть бути повністю дотримані в тестових завданнях, створених ChatGPT. Разом з тим, дослідження виявило труднощі, пов'язані з критерієм надійності, зокрема в контексті неоднозначності варіантів відповідей, згенерованих ШІ.

У статті наголошується на ключовій ролі людського втручання в процеси створення тестових завдань, згенерованих ШІ, та їх вдосконалення. Зокрема зазначається, що детальні та контекстно орієнтовані запити, створені викладачами, є критично важливими для використання максимального потенціалу ChatGPT та усунення його обмежень. З метою оптимізації роботи викладачів іноземної мови під час оцінювання лексичних навичок студентів у дослідженні пропонуються детальні рекомендації щодо вдосконалення процесу створення тестових завдань, створених ChatGPT, наприклад, розробка точних інструкцій, визначення чіткого контексту, призначення ChatGPT конкретних ролей та поетапне вдосконалення результатів. Автори наголошують, що ці стратегії сприяють надійності та ефективності укладання тестових завдань, згенерованих ШІ, та забезпечують їх відповідність освітнім стандартам. Автори підкреслюють важливість інтеграції людського контролю з інструментами ШІ з метою підтримання валідності та корисності мовних тестів.

Це дослідження сприяє ширшій дискусії про інтеграцію ШІ в освіту, демонструючи, як викладачі можуть ефективно використовувати ChatGPT для розробки тестових завдань, враховуючи його переваги та обмеження. Визначено, що майбутні напрями досліджень варто присвятити оцінюванню ефективності інших типів завдань, створених ШІ, вивченню ролі ШІ в автоматизованому оцінюванні та наданні зворотного зв’язку.

**Ключові слова:** штучний інтелект; ChatGPT; розробка завдань для перевірки лексичних навичок; англійська мова як іноземна/англійська мова для спеціальних цілей; тестові завдання, згенеровані ШІ; критерії тестування мовних знань.