

COMPARING AI- AND INSTRUCTOR-DESIGNED ESP TASKS FOR MEDICAL ENGLISH STUDENTS

Ilina Doykova
Medical University “Prof. Dr. Paraskev Stoyanov”

Abstract: *The goal of the study is to explore the degree of equivalence between assessment tasks, created by ESP instructors and AI-generated tasks in the medical field. We compared AI-generated medical texts and types of communicative language tasks, designed for the assessment of medical English competence at level B2 with instructor-generated tasks. We observed the language register and style of the academic texts, as well as the use of formulaic patterns and specific terminology in a summative test. Our findings demonstrate that despite the fact that all AI-generated tasks required further modifications by a human expert, ChatGPT was useful in providing key vocabulary items and task variation in test development. The quality of the generated texts for receptive tasks was rather mechanic and lacking vigour, while the quality of the vocabulary tasks in terms of lexical and syntactic complexity was sufficiently high due to the provided prompts. In ESP practice, ChatGPT can be a useful tool for drafting tasks, vocabulary input, and structural variation. However, the instructor's expertise remains crucial for ensuring authenticity, accuracy, and meeting the assessment goals. Our results indicate that AI is most effective when incorporated into a supervised, human-AI workflow that meets the communicative and disciplinary specificity of ESP assessment.*

Keywords: *English for Medical Purposes; task design; ESP materials development; academic register and style.*

Introduction

Effective ESP teaching and development of assessment materials for medical students incorporate authentic or realistic texts such as case reports and medical histories, clinical guidelines, patient information leaflets, research papers and doctor-patient dialogues to familiarize learners with the discourse of medicine (Hutchinson, Waters 108-109). Designing relevant learning and assessment tasks for this domain can be challenging and time-consuming, as materials must be medically accurate, pedagogically sound, and linguistically appropriate. In this paper we explore the use of ChatGPT to assist the instructor in designing ESP tasks for assessment purposes. We investigate how AI-generated tasks compare to instructor-generated tasks and how well they meet pedagogical objectives. We draw on Biber's multidimensional model of register (Biber 348) as a framework to compare the linguistic characteristics, the lexical and syntactic complexity, and the overall quality of the generated tasks to determine to what extent AI can emulate the style and rigor of instructor-crafted tasks in an English for Medical Purposes course (EMP). The generated range of common task types for assessment of both receptive and

productive skills of second year medical students (following a 60-hour EMP course at level B2) included reading comprehension and listening texts, gap filling, matching, ordering, problem-solving, true/false, multiple-choice questions, case scenarios, lexical collocations and opinion-gap activities. As effective medical communication involves patient-centered language, rapport-building, and shared understanding, we additionally examined the “empathy dimension” in doctor-patient interaction tasks (Silverman 119–148, Dvořáčková 35) or whether AI can modulate the tone of its output and how this compares to good practices in medical communication.

ChatGPT has been trained on massive corpora of texts and is register-aware, i.e. it can mimic various registers and styles of writing. Such proficiency suggests that we could draft EMP tasks that feel authentic to the style of professional medical communication. Kung et al. (2023) found that ChatGPT was able to answer United States Medical Licensing Exam (USMLE) questions at a passing level, demonstrating not only linguistic fluency but also significant medical knowledge (Kung 1). Moreover, instructors may use ChatGPT to make quiz questions, summarize readings, or as an online tutor (Kasneci 5). Other potential benefits include saving time for educators and providing instant practice materials for learners, customization and adaptability of the assessment questions, variety and wide coverage of content as well as the possibility for regenerating items until they match pedagogical needs (Education week). However, multiple concerns regarding the application of AI models for educational purposes have already been raised – the factual accuracy is not guaranteed as the model might produce incorrect statements that sound plausible (Huang 2, Ložić 104, Zhai 9), the text quality can sometimes be generic or lack the depth and nuance that a subject matter expert would provide (Schroeder 7–9) as well as the need for a careful contextual adaptation and human judgment remain irreplaceable (Kuznetsov 247–258).

Consequently, the objectives of our research are the following: (1) to create a set of medical ESP tasks using ChatGPT; (2) to compare the linguistic features and style of these AI-generated tasks to those of instructor-written tasks; and (3) to assess the pedagogical quality of the tasks in terms of clarity, level of difficulty, and authenticity through expert review. We expect ChatGPT to be especially good at producing grammatically correct, content-rich, accurate and lexically appropriate test tasks as part of a standardized test in EMP at level B2. We also expect that with well-crafted prompts, the AI can adjust the level of formality in terms of register and tone to an extend where minimal or no editing is necessary by an instructor.

Methodology

Based on past exam papers, we chose a set of 10 tasks that are commonly used in EMP courses to assess both receptive and productive skills as follows:

1. Reading comprehension (~ 500 words) with 10 open or multiple-choice questions (MCQs) or True/False statements
2. Listening comprehension (~ 500 words) with 10 open or MCQs
3. Use of English (vocabulary gap-filling)
4. Use of collocations
5. Matching terms and definitions
6. Reformulation tasks
7. Case scenarios
8. Role-play (doctor-patient conversations)
9. Grammar-in-use (focusing on forms common in medical texts such as modal verbs, conditionals, passive voice, relative clauses, quantifiers, reported speech, etc.)
10. Short answer questions (open questions)

Task-based language assessment aligns with OSCE-style assessment principles (Douglas 92–128, Norris 578–594) and is particularly relevant in medical education as it allows for simultaneous assessment of the linguistic accuracy, appropriateness (medical terminology, grammar, register) and the communicative competence (clarity, coherence, empathy, intercultural awareness) of medical students.

By evaluating the conformity of AI-generated assessment materials to the anticipated linguistic standards of their respective contexts (ESP and EMP), test developers can guarantee both authenticity and instructional effectiveness. Therefore, we performed a functional profiling of the generated texts, and applied the Biber's multidimensional model (MD) to identify stylistic inconsistencies, genre blending, or over-generalization – tendencies frequently seen in large language model (LLM) output. Unlike surface-level metrics such as lexical diversity or readability scores, used in corpus linguistics, the MD model offers a functionally oriented method for evaluating the register appropriateness of the AI-generated texts in educational contexts by examining co-occurring linguistic features that cluster together to reflect underlying communicative functions (Biber 125–130; Conrad, Biber 13–42).

The aspects of English register variation are briefly summarized below (Biber 348).

Dimension 1: Informational Production versus Involved

This dimension compares content-focused, referential, and information-dense (informational) language with interactional, affective, and instantaneous (involved) language. While informational texts (e.g., academic prose) utilize more nouns, sophisticated phrases, and a wide range of vocabulary to convey facts, involved conversation typically uses first- and second-person pronouns, questions, contractions, and other indicators of interpersonal engagement. Academic medical texts tend to be objective and evidence-based, and thus more informational in nature.

Dimension 2: Production that is narrative as opposed to non-narrative

This feature sets expository speech apart from storytelling. Present tense or timeless statements are preferred in non-narrative writings (e.g., research articles or educational resources, instructions, etc.), while a temporal event sequence (e.g., a case history) indicates a high narrative score.

Dimension 3: Situation-dependent versus explicit reference

Technical and academic writing is known for high explicitness. Noun phrases, qualifiers, and relative clauses are used in explicit texts to make sure all referents are precise and self-contained. On the other hand, situation-dependent texts use pronouns and adverbs that are specific to the immediate context, presuming that the reader or listener is aware of the circumstances (informal speech or in-person instructions often rely on context).

Dimension 4: Direct Persuasion Expression

Suasive verbs (e.g., recommend, suggest), modal verbs of necessity and prediction, infinitives to express intentions or suggestions, and conditional clauses are examples of elements found in persuasive texts. A neutral, solely informative tone is indicated by a low score on this dimension, whereas an argumentative or advisory tone is indicated by a high score.

Dimension 5: Information that is abstract as opposed to non-abstract

This dimension reflects the degree of abstraction and impersonal manner of a text. Academic writing is characterized by texts that use more nominalizations, passive constructions, and impersonal structures to convey abstract concepts (theoretical, hypothetical, conceptual, etc.). “Non-abstract” texts, on the other hand, are more detailed and intimate, use daily language, human agents, and the active voice.

We would add an extra “empathy component” to the five dimensions above, which is pertinent to healthcare communication. Using inclusive first-person plural pronouns, hedging strategies to avoid making direct or absolute assertions, expressing understanding, and using a kind, consoling tone are all examples of the empathic communication style in doctor-patient interactions.

Building trust in medical encounters requires such linguistic cues to demonstrate care and respect for the patient. We evaluated ChatGPT's ability to incorporate sympathetic language into clinical communication tasks thus preventing an impersonal or detached tone in the output.

Results

We formulated targeted prompts to generate the task texts, which were then qualitatively analyzed for linguistic features corresponding to Biber's five dimensions. The AI-generated outputs were independently reviewed by two EMP instructors to evaluate factual accuracy, clarity, and overall language appropriateness as well as to identify occurrences of salient linguistic features (e.g., pronoun versus noun usage, verb tense choices, modality). In the following section we provide sample AI-generated tasks and the prompts used:

Task 1. Write a listening comprehension text (500 words) on anosmia. Use medically accurate language at level B2. Define the term, give common causes, possible complications, and relevant context (e.g., age factors). Formulate 10 comprehension questions and their short answers according to the text:

Anosmia is the inability to detect an odor, whether it comes from pies fresh from the oven or smelly socks piled in a corner. It's usually a temporary side effect from a cold or sinus infection that goes away when the cold or sinus infection clears. In addition, people's sense of smell fades as they age, so people age 50 and older may have long-lasting anosmia. In some cases, anosmia may be a symptom of other, more serious medical issues like diabetes or a traumatic brain injury.

How does the sense of smell typically work? The process starts with substances that smell and give off tiny molecules. When one inhales, the molecules glide into the nose and land on a tiny patch of tissue high inside it....

1. What is anosmia?
2. What usually causes temporary anosmia?
3. What age group may experience long-lasting anosmia? ...

Task 2. Produce a reading task with 10 comprehension questions about eHealth. Include 10 comprehension questions and their short answers.

eHealth – the use of information and communication technologies to support health and healthcare – has gained significant momentum in recent years.

From mobile apps that track chronic conditions to virtual mental health consultations, eHealth has redefined how care can be accessed, delivered, and monitored. In an increasingly connected world, this digital shift offers the potential to improve health systems' efficiency, reach, and responsiveness. However, the reality of eHealth implementation is more complex, raising questions about equity, infrastructure, and governance.

As healthcare systems around the world seek to recover from the long-term effects of the COVID-19 pandemic, many are turning to digital tools as a way to build resilience....

1. What is eHealth?
2. Name one way in which eHealth supports healthcare delivery.
3. Which global event accelerated interest in eHealth?
4. What major issue arises from incompatible health platforms? ...

Task 3. Make a list of 10 medical terms and their lay term equivalents.

Hypertension – High blood pressure	Dyspnea – Shortness of breath
Myocardial infarction – Heart attack	Tachycardia – Rapid heartbeat
Cerebrovascular accident – Stroke	Bradycardia – Slow heartbeat
Neoplasm – Tumor	Analgesic – Pain reliever
Malignant – Cancerous	Edema – Swelling

Task 4. Design a matching term and definitions task with 7 terms from general medicine.

1. Intravenous	A. Pertaining to the skin
2. Hemorrhage	B. Sudden abnormal surge of electrical activity in the brain, resulting in involuntary body movements or behaviors
3. Cutaneous	C. Three layers of membranes that cover the brain and spinal cord
4. Malignant	D. Rapid loss of blood, as in bleeding
5. Seizure	E. Disease in which excessive uric acid in the blood causes sodium urate crystals to deposit in the joints, especially the great toe, producing arthritis
6. Meninges	F. Pertaining to within a vein
7. Gout	G. Tending to become progressively worse, possibly resulting in death

Task 5. Form 10 sentences at language level B2. Use linking words for contrast to compare animal and plant cells.

1. Plant cells have a rigid cell wall, *whereas* animal cells do not, having only a flexible cell membrane.

2. Plant cells contain chloroplasts for photosynthesis, *while* animal cells lack chloroplasts and instead rely on consuming organic matter for energy.
3. Plant cells have a large central vacuole, *in contrast*, animal cells have smaller, more numerous vacuoles with varied functions....

Task 6. Gap fill. Use the key words "incidence," "irritant," "exacerbation," and "dosage" in medical sentences.

1. The *incidence* of asthma has been steadily increasing over the past decade, particularly in urban areas.
2. Exposure to cigarette smoke can act as an *irritant*, worsening symptoms for individuals with chronic bronchitis.
3. During the winter months, many patients experience an *exacerbation* of their COPD symptoms due to the cold air.
4. It is crucial to follow the prescribed *dosage* of antibiotics to ensure the infection is completely eradicated and to prevent resistance.

Task 7. Write 10 single best answer multiple choice questions with four options based on the text.

1. What is the primary function of antibiotics?
 - A) To kill or stop the reproduction of bacteria
 - B) To inhibit viruses from reproducing
 - C) To directly kill fungi
 - D) To increase the body's natural defenses
2. Which of the following conditions can antibiotics cure when used properly?
 - A) Influenza
 - B) Common cold
 - C) Strep throat
 - D) Athlete's foot
3. Why is it important to take an antibiotic as directed, even after symptoms disappear?
 - A) To avoid side effects
 - B) To prevent the development of resistant bacteria
 - C) To ensure the body's natural defenses are enhanced
 - D) To save costs on future medications ...

Task 8. Writing instructions. Explain to a patient in the pharmacy how to take fiber supplements (a laxative).

1. It's best to take your fiber supplement with a meal. This can help you remember to take it regularly and can also reduce any potential stomach upset.
2. If you're new to fiber supplements, start with a small dose. This helps your body adjust and can prevent bloating or gas. You can gradually increase the dose as your body gets used to it.
3. Always take your fiber supplement with a full glass of water. This is crucial because fiber needs water to work properly. Without enough water, the supplement can actually make constipation worse.
4. Drink plenty of fluids throughout the day. Aim for at least 8 glasses of water daily to keep things moving smoothly in your digestive system.
5. Read the label on your fiber supplement carefully. Different products have different dosages and instructions.
6. Fiber supplements can take a little time to start working. You might not see immediate results, but with consistent use, you should notice an improvement in your bowel movements within a few days to a week.

Task 9. Write a short case study: a patient suffering from allergic rhinitis

John, a 15-year-old high school student, has been experiencing sneezing and a runny nose for the past 3 weeks. His symptoms include nasal congestion, nasal itching, and watery, irritant eyes. He also reports a thick discharge from his eyes. John has a previous history of asthma and eczema. His symptoms worsen in the morning and on windy days. He is currently not taking any medication for his symptoms.

Task: *Read the case and write follow-up questions to clarify onset, triggers, and the severity of symptoms.*

One specific task was designed to examine if ChatGPT can implicitly follow the pragmatic norms of medicine where doctors are expected to cushion information to patients and present it in a more empathetic manner (Task 10).

Task 10. Modify the statements to show empathy:

1. You will have an upset tummy.
2. This medication makes you drowsy.
3. You'll develop acne.

Here are more empathetic versions of those statements:

1. You might experience some stomach discomfort, but we'll do our best to manage it.
2. This medication may make you feel drowsy, so it's important to rest when needed.

3. You could develop some acne, but we'll monitor it closely and address any concerns you have.

The AI-generated texts largely conformed to the informational, expository register, as expected for academic or professional medical content. The AI outputs were also rich in specialised vocabulary, grammatically accurate and distinctive of the medical register and style. For example, no first or second person pronouns were found in the listening passage where a medical condition was presented in a general sense. Most of the sentences in that passage were multi-clausal, packed with detailed information (expressed with nominalizations), with an average sentence length of about 20 words. In a similar vein, the reading comprehension section focused on facts rather than subjective experience or opinion by using passive and third-person expressions. The non-narrative and informative nature of the generated texts indicates that ChatGPT ensures the objective, factual tone for these types of requests. Nevertheless, the model did generate a narrative when asked for a case scenario; it was simple and clinical in style, similar to a brief case presentation seen in a medical textbook. In general, the AI-generated content was explicit in its references and self-contained. All necessary information to answer the questions was included in the text, with no unexplained pronouns or references that would confuse a student. By default, the AI outputs were descriptive or instructive without using persuasive language. For instance, the reading and listening texts presented facts and explanations without any imperative sentences or direct appeals to the reader. We did not find rhetorical questions, nor emphatic phrases or loaded emotional adjectives in the reading and listening texts in line with academic neutrality.

Task 8 was very specific in providing step-by-step guidance in imperative sentences, listing tangible actions (*take with water, drink fluids, read the label*). The instructions were clear, logically ordered, and complete. This output demonstrates that ChatGPT can produce non-abstract, highly specific information in appropriate format and style.

On the other hand, tasks that simulate interaction (e.g. a doctor-patient dialogue) showed more involved features, though still relatively formal. The sentences after empathetic rephrasing (Task 10) included first- and second-person pronouns that immediately engage the interlocutor, reflecting an interactive, interpersonal tone and affective engagement (*we'll do our best to manage it*). This demonstrates that ChatGPT may adapt to persuasiveness and empathy by transforming authoritative comments into collaborative and reassuring ones for pragmatic purposes.

The linguistic analysis indicates that ChatGPT's outputs conform to the stylistic standards of academic and professional medical writing. The default tone is formal, objective, and information-rich, necessitating minimal editing.

When an alternative tone was required (empathetic, narrative, etc.), the model could shift registers with appropriate prompting. Nonetheless, the style exhibited excessive consistency, occasionally rendering it mechanically rigid or formulaic. The reading passage resembled a condensed encyclopedia entry: accurate and lucid yet lacking in engagement or originality, devoid of the “human touch” or unique voice that an author might inject. This is likely due to the model's tendency to average across numerous texts, resulting in a cautious, neutral approach.

Regarding content accuracy, we found the AI-generated texts to be accurate in the information they presented, especially for well-established medical facts. In the patient case (Task 9), the symptoms listed (sneezing, runny nose, itchy eyes, etc.) are consistent with allergic rhinitis, which is likely what the case was pointing at; nothing contradictory was present. The answers to MCQs were correct according to the text provided and the standard medical knowledge. This reliability was reassuring, but it can also be attributed to the fact that we guided the content with our prompts and selected topics that were not obscure. In terms of more complex or controversial medical topics, an outdated or incorrect information is likely to be produced.

In terms of quality of content both the reading and listening passages covered the topic adequately but might not spark students' interest. The generated text exhibited extremely high informational density (long noun phrases, numerous nominalizations, complex sentences), was entirely non-narrative, and highly abstract. The text was presented as a block of information with no rhetorical variety (e.g., no events, timelines, personal stories or relational markers). In comprehension tasks, B2 learners benefit from a mix of information and “involved” features (e.g., examples, paraphrasing and clarifying remarks, etc.) that reduce the processing load. Clarity is beneficial; yet, the absence of related cues renders the text static and disconnected. The uniform tone may hinder the identification of significant information due to the absence of prosodic or rhetorical indicators that emphasize the importance of specific issues. Thus, despite the fact that the generated comprehension questions matched the typical exam style, the suggested reading and listening texts were considered not suitable for inclusion in a test format as they lacked the authenticity and originality of human writing in terms of syntax structure, discourse markers use, narrative elements and specific detail. This observation reinforces the perspective that ChatGPT is best utilised as a collaborative tool for educators, providing draft material that benefits from subsequent human refinement and contextualisation.

However, the vocabulary and grammar exercises were well constructed, straightforward, contextually relevant and analogous to those an instructor might produce. The dialogue (doctor-patient conversation) that was generated in one trial was polite and medically correct, but felt scripted as it

followed a pattern of greeting, question, response, reassurance, etc., perfectly (as per the SPIKES protocol), which might not reflect all the spontaneity of real speech.

In summary, the post-editing modifications, which the AI outputs required, were minor – overly long sentences were broken into shorter ones for easier reading; nominalizations were reduced; some listening and reading questions were revised for clarity and simplicity; key words and intentional vocabulary were used to formulate the prompts (Task 6). Thus, the most time-intensive stage in preparing AI-assisted test tasks was the selection of high-quality, authentic texts by the instructors to maintain engagement and bring about a dynamic, context-rich nature that a piece of text created from real experience might have. Once an appropriate text was identified, AI efficiently generated comprehension questions, extracted key vocabulary, and proposed supplementary language tasks. This automation significantly reduced drafting time, enabling educators to focus on reviewing and aligning materials with assessment goals.

Discussion

This study adopts an exploratory, qualitative approach in the comparison between AI-generated and human-authored tasks. Quantitative measurement of the five dimension scores was not undertaken, but could be pursued as a subsequent phase through the application of computational analysis tools to the texts. Our results highlight both the promise and the limitations of using AI to assist in designing EMP tasks for assessment purposes. From a register perspective, AI-generated texts showed a high degree of alignment with the target genre (informational, formal, technical). One key finding is that AI-generated tasks can reach a level of difficulty and complexity comparable to instructor-generated tasks, especially when prompts are carefully constructed. In our case, the vocabulary load and syntactic complexity of the reading and listening passages were more complex and abstract than those found in authentic materials for that proficiency level.

Our findings indicate that AI may exhibit pragmatically suitable methods for addressing sensitive subjects (e.g., poor prognosis), which has implications for using AI during instruction, not just in material preparation to build awareness of tone. We consider the ChatGPT's capacity to integrate empathy and modify persuasion upon request as a valuable asset in communication skills training.

From an assessment standpoint, the dual capability of AI as item writer and provider of answers could be used to quickly produce practice quizzes with answer explanations for students. This capability raises issues about academic integrity, as AI's ability to solve generated assignments suggests that

students might likewise use AI to acquire answers. In our case, the assignments were performed under supervision; however, for take-home or unattended/online tests, it is imperative to acknowledge that an astute student may simply consult ChatGPT for the same inquiries.

The comparisons with instructor-generated tasks in this study highlighted the importance of human-written texts for inclusion in comprehension tasks. Most importantly, in terms of content selection, experienced instructors – particularly those who are familiar with the terminology, collocations, and specialised vocabulary of an EMP or ESP course – are best positioned to formulate effective prompts for AI-assisted task generation. Their ability to distinguish between linguistically central and peripheral vocabulary allows them to create tasks that are in line with course objectives, reflect real-world language use, and focus on the particular communicative competencies needed in the medical field. Thus, the instructor's role is essential for maintaining the educational relevance and validity of the content and for mitigating any hazards associated with an overdependence on AI-generated materials.

In order to ensure authenticity, many ESP practitioners construct a mini-corpus of pertinent texts from which they extract instructional materials. ChatGPT has essentially done that behind the scenes as it internalized a massive corpus and is now a queryable entity that can produce corpus-like output on demand. This functionality is particularly powerful for testing “keywords in context”, illustrating usage in sentences and terminology-related tasks. We regard this as a major strength, as ChatGPT can simultaneously serve as a thesaurus and an example generator, consolidating functions that would traditionally require multiple separate resources. However, unlike a static corpus, AI may occasionally generate a context that is somewhat inaccurate or less conventional (e.g., drafting mini-scenarios for role-play tasks). While AI can serve as a dynamic corpus-based tool, using it well requires the instructor to have a corpora-informed mindset and expertise – essentially verifying that the usage is indeed common/frequent.

Drawing from our observations, we present several cautionary points for consideration. First, we may not have identified major factual errors in the task outputs as we composed prompts, specific enough to avoid linguistic issues, but research indicates that language models can “hallucinate” or produce incorrect information confidently (AAMC Home Page). Second, we did not explicitly test for bias in our content (e.g., gender/racial bias in medical scenarios). Third, if an instructor becomes too reliant on AI for materials, there is a potential erosion of their own material design skills or a risk of homogenization of tasks. Variety frequently arises from human creativity and insights. Thus, one notable benefit of human involvement is that instructors

have the ability to formulate questions that necessitate personal input or reflection, which AI is not easily capable of generating.

Conclusion

ChatGPT can serve as a corpus-informed text, question, and example generator when utilized judiciously, and as a sophisticated co-creator under the guidance of an instructor. In terms of efficiency, ChatGPT proved useful when guided by clear, specific prompts and produced tasks of comparable difficulty to instructor-designed ones. It can formulate multiple versions of a scenario for practice or quickly update case parameters with new data. Its ability to adjust tone further supports its potential value in ESP/EMP contexts. In contexts where instructors are non-native English speakers, such capabilities are especially useful, enabling the creation of diverse, adaptable materials that enrich ESP courses and support tailored learning experiences across proficiency levels.

At the same time, several limitations must be acknowledged. Our study had a narrowly defined scope, focusing on one proficiency level and a communicative task set for students of Medical English. Larger-scale studies should incorporate quantitative linguistic measures (e.g., multidimensional scoring, complexity indices), alongside with student-based evaluations of difficulty and authenticity. Further work is also needed to examine AI-generated clinical communication, especially its pragmatic and empathetic adequacy compared to authentic dialogues.

Works Cited

Association of American Medical Colleges (AAMC) Home Page. 27 Feb. 2025. *AI in medical education: 5 ways schools are employing new tools*. 30 July 2025.

Biber, Douglas. "On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson." *Text & Talk* 15.3 (1995): 341-370.

Biber, Douglas. *Variation across speech and writing*. Cambridge university press, 1991.

Conrad, Susan, and Douglas Biber. "Multi-dimensional methodology and the dimensions of register variation in English." *Variation in English*. Routledge, 2014. 13-42.

Douglas, Dan. *Assessing languages for specific purposes*. Cambridge university press, 2000.

Dvořáčková, Veronika. "A genre-based approach in ESP classes to teaching clinical communication focusing on breaking bad news to patients." *Discourse and Interaction* 17.1 (2024): 30-50.

Education Week Home Page. 18 June 2025. Ferlazzo L. AI Can Save Teachers Time and Stress. Here's How. 10 Aug. 2025.

Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *ACM Transactions on Information Systems* 43.2 (2025): 1-55.

Hutchinson, Tom, and Alan Waters. *English for specific purposes*. Cambridge university press, 1987.

Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." *Learning and individual differences* 103 (2023): 102274.

Kung, Tiffany H., et al. "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models." *PLoS digital health* 2.2 (2023): e0000198.

Kuznetsov, Andrei N. "AI Perspective of Planning and Realization of Competency Potential of Training in Higher Education." *Integration Strategies of Generative AI in Higher Education*. IGI Global, 2025. 247-258.

Ložić, Edisa, and Benjamin Štular. "Fluent but not factual: A comparative analysis of ChatGPT and other AI chatbots' proficiency and originality in scientific writing for humanities." *Future Internet* 15.10 (2023): 336.

Norris, John M. "Task-based teaching and testing." *The handbook of language teaching* (2009): 578-594.

Schroeder, Hope, et al. "Large Language Models in Qualitative Research: Uses, Tensions, and Intentions." *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025.

Silverman, Jonathan, Suzanne Kurtz, and Juliet Draper. *Skills for communicating with patients*. crc press, 2016.

Zhai, X., et al. (2024). Educational uses and misuses of ChatGPT: A systematic review. *Computers & Education: Artificial Intelligence*, 5, 100196.