Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Special Issue on CALL Vol. 17, No.12. July 2026 Pp.400-417
DOI: https://dx.doi.org/10.24093/awej/call12.23

Homogenization and Epistemic Dependency: How Generative AI Threatens Task-Based Language Assessment in ESP Contexts

Nataliia Avsheniuk
Foreign Systems of Pedagogical and Adult Education Department
Ivan Ziaziun Institute of Pedagogical and Adult Education of the NAES of Ukraine, Kyiv, Ukraine
Email: nataliya.avshenyuk@gmail.com

Nataliia Seminikhyna
Department of Foreign Languages, Faculty of Economics
Taras Shevchenko National University of Kyiv, Ukraine
Corresponding author: nseminikhyna@gmail.com

Olena Lutsenko
Department of Foreign Languages for Natural Sciences Faculties,
Educational and Scientific Institute of Philology,
Taras Shevchenko National University of Kyiv, Ukraine
Email: o.lutsenko@knu.ua

Tetiana Svyrydiuk
Department of Foreign Languages, Faculty of Economics,
Taras Shevchenko National University of Kyiv, Ukraine
Email: tetiana.svyrydiuk@knu.ua

Larysa Ruban
Department of Foreign Languages, Faculty of Economics
Taras Shevchenko National University of Kyiv, Ukraine
Email: larysa.ruban@knu.ua

Received: 03/15/2026 Accepted:07/03/2026 Published: 07/15/2026

Abstract

Task-based language assessment assumes that the learner's performance reflects the learner's own developing communicative competence. Generative AI challenges this assumption by allowing students to delegate task completion to an external system, thereby threatening task authenticity, performance authenticity, and the interpretability of form. The study aims to examine how 40 university English for Specific Purposes teachers at three Ukrainian universities experience the impact of generative artificial intelligence on the task-based language teaching and assessment validity chain. This issue is significant since, in the absence of institutional AI policy, teachers are the primary regulators of AI use in assessed contexts, and their decisions are the primary mechanism through which validity is maintained or compromised, yet remain empirically under-examined. The study employed a qualitative approach, including analytical content analysis of open-ended survey responses and structured classroom observations of 12 teachers (a purposive subsample). The results were five-fold: teachers found artificial intelligence useful for task material generation and formative feedback; teachers had consistent validity concerns for summative assessment; formative-summative asymmetry emerged in different disciplinary clusters; teachers independently indicated that students' output converged towards a single artificial intelligence-modified output; and students would not be able to critically evaluate the artificial intelligence-generated

text. Our findings are interpreted using the Task-Based Language Teaching-Assessment-Artificial Intelligence Nexus, a preliminary heuristic model that identifies three interconnected nodes of validity threat. The study argues that this suggests the need for formal AI policy, and for teaching AI-critical genre literacy in English for Specific Purposes education.
Keywords: Artificial intelligence, assessment validity, English for specific purposes, epistemic dependency, linguistic homogenization, task-based language assessment
Cite as: Avsheniuk, N., Seminikhyna, N., Lutsenko, O., Svyrydiuk, T., & Ruban, L. (2026). Homogenization and Epistemic Dependency: How Generative AI Threatens Task-Based Language Assessment in ESP Contexts. *Arab World English Journal (AWEJ) Special Issue on CALL Vol. 17.* (12)400-417. https://dx.doi.org/10.24093/awej/call12.23

Introduction

The generation of artificial intelligence poses the biggest challenge to language assessment. When students use artificial intelligence to complete assessed tasks, the resulting performance may reflect the system's capabilities rather than the learner's communication skills. This is now more apparent since student submissions are increasingly linguistically complex and standardised, and students are not always able to articulate or justify their answers, or to assess the veracity of their work. Thus, language education is in threat when we integrate artificial intelligence into learning and assessment without compromising on assessment quality or the very conditions for meaningful language development.

This is particularly difficult when task-based language education and evaluation are concerned, because tasks are the primary tool for language development and the primary source of assessment data. The rationale behind this approach is that well-designed communicative activities produce real language use that provides accurate evidence of the learner's development toward a good level of competence. The link becomes problematic when artificial intelligence plays a large part in the task. The task is no longer a learning opportunity, and its result is no longer a reliable measure of the learner's ability. Problems of assessment validity become problems of learning quality.

Such considerations are particularly important in English for Specific Purposes environments. Thus, in those cases, assessment extends beyond language competence to disciplinary communicative competence, that is, the ability to write professional genres (clinical case reports, technical specifications, corporate contracts) and other discipline-specific tasks. This competence is generated through sustained engagement in real communication tasks and professional discourse. Extending task completion to artificial intelligence calls into question the validity of conclusions drawn from actual performance.

In Ukraine, the context of higher education is very relevant to the current investigation of these issues. English for Specific Purposes is an important component of professional degree courses across all discipline areas in the Bologna Process, and disciplinary communication skills are widely known as an important graduate outcome (MES Ukraine, 2024). At the same time, the rapid adoption of artificial intelligence technology has been accompanied by a lack of assessment strategy for higher education. At the time of data collection, whilst international organisations had highlighted the importance of human-centred approaches to artificial intelligence governance in education, and national recommendations had discussed artificial intelligence in general secondary education, there was no comprehensive guidance on assessment in higher education (UNESCO 2023). So, there is no longer explicit guidance from the school, teacher, or professor on how to interpret, regulate, and respond to Artificial Intelligence in the context of assessment.

However, despite growing concern about assessment validity and the use of artificial intelligence, little research has examined how ESP instructors cope with these challenges in their daily teaching practice. This gap is practical and academic. Teachers observe changes in student performance, assess work that may be inaccurate, and develop assessment responses in contexts with very little institutional oversight. Their perspectives thus provide a valuable foundation for understanding how artificial intelligence is shaping the relationship between learning, assessment and validity in English for Specific Purposes education.

The present study examines the perceptions of 40 university professors of English for Specific Purposes from three Ukrainian universities regarding the impact of artificial intelligence on the validity chain in teaching and assessment. The study employed a qualitative approach to analysis of open-ended survey responses and structured classroom observations of a purposive subsample of 12 instructors for four research topics, which were based on three interrelated validity-threat nodes.

RQ1. What do university ESP professors regard as the suitability of AI-generated task materials in terms of genre for their own disciplinary settings, and what criteria do they employ to assess the appropriateness? (Threat to Internal Validity)

RQ2. What is the difference between the usefulness of formative input from AI and perceived validity in summative assessment from ESP teachers' perspective, and what are the explanatory factors? Threat to Validity of the Assessment

RQ3. How often do ESP teachers perceive AI-induced linguistic homogeneity in student output, and how do they define its link to learners' aptitude for critical assessment of AI-generated language? (Threat to Interpretive Validity Construct)

RQ4. What is the ESP teachers' perception of the gap between their professional development requirements connected to AI and the institutional offering, and which digital capabilities do they regard as most urgently needed?

Literature Review

The validity chain in Task-Based Language Assessment (TBLA) is based on three inferential steps: the task elicits performance representative of a real-world communicative context; the learner's performance is representative of authentically developing competence; and the teacher's interpretation of that performance as evidence of the target construct is reasonable (Bachman & Palmer, 2010; Norris, 2009, 2016). Each step needs to support unified construct validity (Messick, 1989). Importantly, this chain of validity is not only psychometric but also educational. TBLT is the engine of acquisition (Ellis, 2003; Long, 2015), and the task performance evaluation is valid precisely because it captures the learning outcome the task was supposed to produce. The validity logic of TBLA is directly derived from TBLT. A validity failure in TBLA is consequently also a pedagogical failure in TBLT: if the performance being assessed has not been carried out through real-task engagement, then neither the assessment nor the learning it was supposed to certify has taken place. TBLA exercises are expected to be authentic, interactive and consequential. In ESP contexts, authenticity entails not only communicative sincerity but also genre-level representativeness (Bygate, 2008). A task that produces general academic writing when the goal is producing medical case reports does not satisfy the authenticity requirement, irrespective of the language complexity (Bhatia, 1993; Swales, 1990).

TBLT helps select tasks in the target task requirements analysis. It is assumed that the communicative need of a task is not just a setting for language usage, but the main mechanism of acquisition (Ellis, 2003; Long, 2015). It is the truth of this acquisitional claim that makes TBLA

valid, because tasks promote learning, and measuring task performance directly reflects the competence those tasks were supposed to generate. The TBLT-TBLA link is not accidental but constitutive: each needs the other for its theoretical coherence. The acquisitional basis of TBLT is interlanguage theory: Corder (1967) established that learner errors are systematic evidence of a developing rule-governed system; Long (1996) argued that negotiation of meaning during task completion, especially when communicative breakdown forces attention to form, is the engine of development. If a student delegates output to AI, the communicative need is avoided: no detection, no productive effort, no individual output that exposes interlanguage character. The work is administratively complete when the acquisitional requirements required by TBLT are not satisfied, and TBLA loses the very evidence foundation on which its validity rests.

One important issue is the various validity implications of using AI in formative vs summative scenarios. If students are evaluatively engaged with AI-generated feedback in formative assessment, rather than passively, it can facilitate iterative draft development without endangering construct validity. However, summative assessment provides significant inferences about learner competence and professional preparation, whereas AI-mediated output consistently contradicts the performance criterion, thereby reflecting the learner's own evolving ability. Valerio (2025) supports this division by warning against replacing traditional teaching methods or direct teacher-student interaction with AI-powered writing tools. This is especially relevant to summative ESP assessment, where the teacher's ability to observe and assess authentic student performance is an irreplaceable validity mechanism.

ESP takes this further by attempting to describe what a genuine performance should look like in specific discourse communities (Dudley-Evans & St. John, 1998; Hyland, 2006). Genre competence, conceived as the ability to enact the move sequences, hedging conventions, evidential norms and authorial positioning patterns of a specific disciplinary community (Swales, 1990; Bhatia, 1993), functions both as the target construct and the criterion for validity in ESP assessment. Large language models, trained mostly on broad-domain literature, recreate a general academic register when asked to develop speciality genres, producing output that is seemingly reasonable but genre-inappropriate in ways that specialist genre specialists can identify (Hyland, 2019). This effect has been termed AI-induced linguistic homogeneity, a systematic conformity of student output to the probabilistic average of LLM output, which may threaten the intellectual distinctiveness and linguistic variety in ESP assessment. This homogenization is reinforced by epistemic reliance, a state in which pupils are not sufficiently genre-literate to critically assess AI-produced texts. Critical AI literacy is probably among the least developed AI capabilities in colleges (Ng et al., 2021). ESP learners with lower competence are differentially exposed, since the very domain expertise needed to recognise input inappropriateness is still being acquired (Long & Magerko, 2020). Research supports authentic affordances of AI as a pedagogical resource rather than an output generator, supporting task material design, scaled formative feedback, and language complexity development among higher-literacy learners (Li et al., 2025; Avsheniuk et al., 2025), but the success of implementation depends on teacher AI literacy and institutional design capacity (Bhandari et al., 2025). Teachers' views are important moderators of technology integration results (Ertmer, 2005; Tondeur et al., 2012). However, the current corpus of AI-in-education studies has been mainly focused on student outcomes and general EFL contexts, resulting in two significant lacunae. First, the perspective of ESP teachers, whose dual expertise in language pedagogy and disciplinary discourse renders their perceptions uniquely informative, is largely absent. Second, the framing has been mainly affordance-centric rather than validity-centric, conceptualising AI as

a tool to be optimised rather than a threat to the inferential chain on which assessment legitimacy rests. What is lacking in empirical studies is an understanding of how practical ESP teachers perceive and respond to these overlapping validity challenges.

Generative AI poses a particular epistemic duty for teacher educators: rather than teaching teacher candidates how to use AI tools, teacher educators are required to encourage the critical pedagogical decision about when the use of AI advances the goals of language learning and when it undermines them (Case et al., 2025). The five-dimensional AI literacy framework, consisting of AI foundations, AI ethics, a human-centred mindset, AI pedagogy and AI for professional development, covers the full scope of what ESP teachers need to manage AI integration responsibly, and it serves as a benchmark against which the current professional development provision can be evaluated (Liu,2025).

ESP instructors are especially in a challenging professional position as practitioners in the intersection of language pedagogy and discipline knowledge (Dudley-Evans & St. John, 1998). It is exactly this dual skill – knowledge of language-pedagogical principles and acquaintance with disciplinary norms of discourse – that is needed to identify AI-induced genre failure and to discern true student voice from AI-inflected uniformity. It is also, perhaps, specifically what is of concern when AI is framed as a replacement for rather than a complement to experienced professional judgment. It is both empirically urgent and philosophically important to know how ESP teachers perceive, respond to, and seek to address the problems of homogeneity and epistemic reliance.

Method

This study employed a qualitative design that integrated two complementary elements: systematic content analysis of open-ended survey responses using a theoretically grounded codebook, with reliability assessed via the Miles and Huberman (1994) percentage-agreement formula; and structured classroom observation of a 12-teacher subsample. The triangulation process involved systematically comparing and synthesising data from the two strands: teacher-reported perceptions gathered through survey responses and researcher-recorded observations of classroom behaviour. This approach enabled the identification of convergences and divergences between what teachers reported and what was directly observed, thereby enhancing the internal validity and robustness of the findings. By drawing on both self-reported and observational data, the study achieved a more comprehensive and precise interpretation of the validity threats under investigation.

Participants

The study involved 40 university ESP teachers from the foreign languages departments of three Ukrainian universities: a technical university, a medical university, and a university of law and economics. Participants were selected through purposive sampling based on two criteria: active ESP teaching practice at the university level and prior experience with generative AI tools in a teaching or assessment context. This criterion-based selection ensured that participants could speak from direct professional experience rather than hypothetical familiarity, which is essential for the credibility of qualitative data in a study focused on practitioner perception. The survey was administered in English without translation, as all participants were practising ESP teachers whose professional work is conducted in English.

Research Instruments

The study employed two primary instruments: an open-ended survey and a structured classroom observation protocol. The survey instrument consisted of open-ended questions designed to elicit detailed practitioner accounts of AI's impact across four thematic areas: task design affordances of AI-generated materials, AI's differential role in formative versus summative assessment, observed linguistic homogenization and epistemic dependency in student output, and perceived digital competency development needs. Initial items were adapted from validated instruments in AI integration research and task-based language assessment scholarship to ensure conceptual alignment with the study's four research questions. Items were then subjected to expert content validity review using a five-panel Content Validity Index procedure, in which five specialists in ESP pedagogy, language assessment, and educational technology independently rated each item for relevance, clarity, and representativeness on a four-point scale. Pilot responses were analysed to identify ambiguous items, inconsistent interpretations, and response gaps, and the instrument was finalised accordingly before the main data collection.

The classroom observation protocol consisted of 24 rating items organised on a four-point frequency scale across three behavioral dimensions corresponding to the study's three validity threat nodes: task design behaviors, capturing how teachers selected, modified, or rejected AI-generated materials in live instruction; assessment performance activity, capturing homogenization markers across student presentations and students' inability to respond to follow-up questions about their own content; and student-AI interaction performance, taking observable patterns of AI reliance during in-class task completion.

Research Procedures

The questionnaire was administered electronically (Qualtrics) over six weeks (March–April 2025). Of 52 teachers invited, 40 completed the full instrument (76.9% response rate). The researchers observed the 12-teacher subsample across one complete ESP unit each (42 sessions total). A structured protocol of 24 rating items (4-point frequency scale) organised observations into three behavioural dimensions corresponding to the three validity-threat nodes: task design behaviours, assessment performance behaviours (homogenization markers, inability to respond to follow-up questions), and student-AI interaction behaviours. A second observer attended 6 of the 42 sessions; inter-observer reliability averaged 84.7% (Miles & Huberman, 1994), exceeding the recommended 80% threshold.

Open-ended responses were analysed through systematic content analysis (Schreier, 2012), using a codebook explicitly constructed from the validity chain framework to ensure theoretical alignment. The coding process involved two independent coders, each assigning codes to the full set of responses according to the predefined categories and subcategories in the codebook. After initial, independent coding, the level of agreement between coders was calculated using the percentage agreement formula described by Miles and Huberman (1994). Only categories that met or exceeded an 80% agreement threshold were advanced to the consensus resolution stage, where disagreements were discussed and definitions further refined as needed. Across all categories, final inter-coder reliability was 85.1%, reflecting strong agreement.

Field notes were subjected to content analysis using the same codebook, with the addition of two observation-specific sub-categories: Observable Homogenization Indicators, such as repeated use of identical phrases across multiple student presentations or consistent structural similarities in written submissions, and Student AI-Dependency Behaviours, including students' inability to answer follow-up questions about their own submitted work or reliance on statements

indicating unfamiliarity with presented content. Findings across the two strands were integrated at the interpretation stage through systematic comparison of content analysis categories with these specified observational indicators, enabling convergent or divergent triangulation.

Results

Category 1: Pedagogical Affordances

Systematic content analysis of open-ended survey responses and structured classroom observation data produced five thematic categories, each comprising two sub-categories (Table 2). Three categories map directly onto the study's validity threat nodes: Category 1 (Pedagogical Affordances), comprising task material generation and personalized feedback sub-categories, addresses the task design node; Category 2 (Summative Assessment and Validity Concerns), comprising construct validity and task resistance sub-categories, addresses the assessment validity node; and Category 4 (Linguistic Homogenization and Epistemic Dependency), comprising homogenization and epistemic dependency sub-categories, addresses the construct interpretation node. Two categories function as cross-cutting enabling conditions operating across all three nodes: Category 3 (Teacher Professional Identity), comprising AI as professional supporter and AI as professional threat sub-categories, captures the dispositional preconditions that determine whether teachers enact adjusted pedagogical responses; and Category 5 (Institutional Support Deficits and Digital Competency Needs), comprising policy absence and digital competency gaps sub-categories, captures the structural conditions that determine whether those responses can be institutionally supported and sustained. Table 2 presents the full category and subcategory structure, along with coded unit frequencies across disciplinary groups.

The most frequently coded category (n = 38, 95%) concerned AI's practical efficiency for task design and formative feedback. The dominant sub-category, task material generation (n = 34), reflected teachers' appreciation of AI as a resource for producing discipline-specific inputs where authentic professional texts were scarce: one participant described AI as:

"A practical way to create tasks that actually sound like real professional communication, not just textbook exercises." (Teacher 4)

The personalised feedback sub-category (n = 28) captured value attributed to AI's capacity to deliver immediate, individualised responses to student drafts at scale. Medical ESP and Legal ESP participants were over-represented across the two sub-categories (18 of 25 coded units), consistent with patterns documented in higher-genre-specificity disciplinary contexts. Observation corroborated this picture: AI-generated materials were used as task inputs in 9 of 12 classrooms, and teachers had visibly annotated or modified the output before use in 7 of those 9 cases.

Category 2: Summative Assessment and Validity Concerns

The second most frequently coded category (n = 35, 87.5%) concerned AI's threat to the validity of summative assessment. The construct validity sub-category (n = 31) captured teachers' difficulty in interpreting AI-assisted student work as evidence of genuine communicative competence, a concern several participants articulated in terms that closely parallel the differential validity logic of Bachman and Palmer (2010) without invoking formal assessment theory. Teacher responses captured this recognition with directness:

"I can see the student has produced something fluent and well-structured, but when I ask them to explain their argument, there is nothing there. The text is not theirs." (Teacher 5)

"For formative work, I do not mind if they use AI to get feedback on a draft. But for final assessment, I need to know what the student can actually do, not what the machine can do for them." (Teacher 6)

"I started asking students to defend their written submissions orally. The gap between what they submitted and what they could say out loud told me everything I needed to know." (Teacher 7)

The task resistance sub-category (n = 22) documented emerging pedagogical responses: oral defence requirements, in-class writing conditions, and process portfolios. Observation provided direct behavioural evidence for this category's central claim: in 6 of 12 observed classes, teachers used follow-up questioning during student presentations, and in 4 of those 6 cases, students were visibly unable to answer questions about content drawn from their own slides. This performance-competence gap was replicated across all three disciplinary clusters.

Category 3: Teacher Professional Identity

Twenty-nine participants (72.5%) reflected on AI's implications for their professional role, producing a finding that is theoretically significant not as a standalone validity concern but as a cross-cutting enabling condition that determines whether teachers have the professional investment to recognise and respond to Nexus threats at all. The category was divided almost equally between AI as a professional enabler (n = 15), freeing teachers from routine feedback to focus on higher-order facilitation and genre evaluation, and AI as a professional threat (n = 14), in which infringement on design and feedback roles was perceived as deskilling. Teacher comments illustrate both orientations with clarity:

"AI handles the mechanical feedback so I can focus on what I actually do best — helping students understand why a particular genre move works in their field and not just whether their grammar is correct." (Teacher 8)

"When AI can generate a better task than I can in thirty seconds, I start to wonder what exactly my role is anymore. It does not feel like a tool. It feels like a replacement." (Teacher 9)

"I use AI as a starting point, not an endpoint. It gives me something to react to, to improve, to make discipline-specific. That is still my expertise — the AI just saves me time on the first draft." (Teacher 10)

"I stopped using AI entirely in my classes because I could not trust what it was producing, and I did not want my students to think that kind of generic output was acceptable professional writing." (Teacher 11)

This near-even split carries a direct implication for the Nexus model: teachers who experience AI as a professional enabler are more likely to engage critically with AI-generated materials at the task design node and to invest in the genre evaluation competencies the Nexus requires; teachers who experience it as a threat may disengage entirely, defaulting to either comprehensive prohibition or uncritical acceptance, neither of which constitutes the adjusted pedagogical response the model specifies. Professional orientation toward AI is therefore not a peripheral finding but a precondition for the pedagogical responses the Nexus proposes, and it is not straightforwardly predicted by career stage or disciplinary context, suggesting that professional development should address identity and agency alongside technical AI literacy.

Category 4: Linguistic Homogenization and Epistemic Dependency

The most theoretically consequential category (n = 33, 82.5%) documented two related phenomena observed independently across disciplinary contexts. The homogenization sub-category (n = 30) captured convergence of student presentations and written work toward a structurally and lexically uniform AI-adjusted speech, a convergence that teachers described as systematically erasing not only interlanguage variation but the individual learner's voice, creative expression, and emerging professional identity — reducing each student to an indistinguishable unit of a probabilistic mean. When all student output is generated by the same probabilistic system, the distinctive style, personal reasoning, and creative linguistic choices through which learners construct and communicate their developing disciplinary identity are extinguished entirely — replaced by a generic, algorithmically averaged register that belongs to no one. Three participant observations illustrate the phenomenon with progressively sharper precision:

"All the presentations sound like they were written by the same person: the same phrases, the same order, no individual voice at all." (Teacher 1)

"Students can produce a technically fluent text but cannot explain what a word means or why a sentence is structured that way." (Teacher 2)

"My students have become editors rather than authors, and they are not even good editors because they cannot tell when the AI is wrong." (Teacher 3)

What these observations collectively suggest is not solely a technical assessment problem but a profound educational crisis: students who outsource language production to AI are not simply submitting inauthentic work. The competence to express one's reasoning, perspective, and identity through the genre conventions of a professional community is precisely what disciplinary language education exists to develop; AI delegation renders that development impossible. The epistemic dependency sub-category (n = 26) captured the underlying competence failure: students unable to detect register inappropriateness, disciplinary inaccuracy, or genre violation in AI-generated text they submitted as their own, and therefore unable to reclaim the creative and intellectual ownership that authentic language production requires.

Structured classroom observation provided convergent evidence from a second data source. Behaviours consistent with homogenization, specifically structural uniformity across at least three presentations within a session combined with shared distinctive lexical items, were recorded in 10 of 12 observed classes (83.3%). Behaviours consistent with epistemic dependency, including an inability to answer follow-up questions about one's own content and the use of phrases suggesting unfamiliarity with one's own text, were observed in 9 of 12 classes (75.0%). The convergence between teacher self-report and researcher observation (82.5% vs. 83.3% for homogenization; 65.0% vs. 75.0% for epistemic dependency) represents the strongest cross-strand triangulation in this dataset. However, it is important to note that these findings are based on perception and classroom observation, rather than direct corpus-level analysis of student language output, which limits the strength of attribution to AI use. Furthermore, students' inability to answer follow-up questions about their own content may have multiple non-AI-related causes, such as test anxiety, insufficient preparation, or general comprehension difficulties. Ascribing such patterns specifically to AI delegation remains tentative without a control group or baseline data for comparison, which this study could not provide.

Category 5: Institutional Support Deficits and Digital Competencies Development Needs

Every participant (100%) produced responses in this category, making it the only one with universal representation. While the universality of this category partly reflects the breadth of the

coding criterion, the specificity of participants' responses goes beyond simple agreement: the policy absence sub-category (n = 38) documented near-universal frustration at the lack of differentiated AI use guidance, with several participants linking this vacuum to the homogenization problem, noting that students defaulted to wholesale AI generation in the absence of institutional policies distinguishing legitimate from illegitimate AI use. Teacher comments captured both the frustration and its consequences:

"There is no policy, no guidance, nothing. Every teacher is doing something different, and students know it. They use AI everywhere because nobody has told them clearly when they cannot." (Teacher 12)

"I do not know what I am allowed to say to students about AI. Can I prohibit it? Can I require them to declare it? I am making this up as I go along, and I should not have to." (Teacher 13)

"The students who use AI most are not the weakest students — they are the ones who have figured out that nobody is checking. That is an institutional failure, not a student failure." (Teacher 14)

The development of digital competencies gaps sub-category (n = 33) specified the most-needed competencies: AI-critical genre evaluation (n = 29), process-oriented assessment design (n = 24), and learner AI literacy instruction (n = 19).

The findings directly address the four research questions posed in this study, illustrating in detail how university ESP teachers in Ukrainian higher education perceive the impact of generative AI on the task-based language teaching and assessment validity chain. With respect to RQ1, teachers recognised the practical utility of AI-generated materials, especially for producing discipline-relevant inputs in fields with limited authentic resources. However, they consistently reported the need to modify these materials before classroom use to ensure their authenticity in the target genre, confirming both AI's affordances and its limitations within the ESP context.

Discussion

The findings revealed how university ESP teachers in Ukrainian higher education perceive generative AI's impact on the task-based language teaching and assessment validity chain across four dimensions. Teachers perceived AI-generated materials as practically useful but genre-unreliable, consistently modifying AI output before classroom use rather than deploying it unmediated (RQ1). They drew a theoretically accurate distinction between formative contexts, in which AI affordances were embraced with confidence, and summative contexts, in which AI-mediated production was perceived as decoupling performance from competence (RQ2). Independently and across all disciplinary clusters, they reported convergence of student output toward a uniform AI-adjusted register alongside students' inability to evaluate or account for the language they had submitted — a pattern whose causal significance extends beyond assessment to the learning conditions themselves (RQ3). Simultaneously, they identified the absence of institutional policy and professional development deficits as the structural condition underlying all other findings, leaving individual teacher judgment as the sole regulatory instrument in cases where a coordinated institutional response is required (RQ4).

Participants most enthusiastic about the task-generation function of AI were those in Medical and Legal ESP, the disciplines with the strictest genre conventions, but also the most concerned about disciplinary reliability. Participants observed that AI-generated medical case discussion reproduces the structural moves but not the evidential conventions and norms of clinical

reasoning. This is exactly the criteria of authenticity employed by Samuda and Bygate (Samuda & Bygate, 2008): a task that elicits genre-appropriate surface structure while violating underlying rhetorical rules does not satisfy TBLA validity standards. AI may be a drafting resource for instructors to analyse and edit for genre appropriateness in TBLT task design, but not an unmediated problem generator. These findings show that professional development of the AI-critical genre evaluation is important and likely to be useful.

Perhaps the most conceptually consistent finding in this dataset is the formative-summative asymmetry in participants' attitudes: formative assessment is significantly more positively evaluated than summative assessment. This is immediately mapped onto the validity chain analysis shown above. Formative assessment provides low-stakes feedback to guide learning, rather than certifying competency. If AI feedback helps a student improve a draft, the validity of that improvement is not contingent on the student having developed the draft independently. The feedback does what it is supposed to do, no matter how it was induced (Meniado, 2023).

In contrast, summative assessment produces major conclusions that are documented, published, and utilised to make judgments regarding learner development, certification, and professional preparation (Norris, 2016). For such inferences to be true, the performance on which they are based must be the authentic output of the learner's own ability, a criterion that AI-mediated creation routinely fails. The participants' replies revealed an intuitive yet theoretically sound understanding of the need for unequal validity. More experienced instructors were more concerned with the validity of summative assessment. One interpretation, consistent with Ertmer's (2005) second-order barrier framework, is that greater professional investment in established assessment standards increases sensitivity to validity threats. Another is that teachers' longer time in the profession means stronger prior commitments to existing assessment formats that AI disrupts. Longitudinal and interview-based approaches would be needed to disentangle them.

The practical implication is that policy approaches to formative and summative AI use must differ: formative use can be allowed and even fostered when students treat AI feedback evaluatively rather than passively; summative use must be either prohibited with process-evidence requirements or a move to performance-under-observation constructs that AI cannot replace. (Norris, 2016; Samuda & Bygate, 2008).

The most theoretically significant findings came from the construct interpretation node, where qualitative data showed that participants had, independently and across disciplinary settings, observed the systematic convergence of student output toward a consistent AI-inflected style. This phenomenon, AI-induced linguistic homogenization, is not just a psychometric problem, but a profound educational crisis: when every student's presentation sounds the same – shaped by the same probabilistic tendencies of the same model – the classroom loses the individual voices, idiosyncratic reasoning, and emergent professional identities that make disciplinary development visible, assessable, and real. Students producing output from probabilistic systems are denied the fundamental battle that creates real disciplinary identity: the fruitful confrontation of the learner's emerging mind with the needs of a particular professional genre. AI eliminates contact, substitutes frictionless delegation, and leaves no trace of the individual behind. This is theorised in the literature review as a structural consequence of LLM architecture (Bender et al., 2021; Casal & Kessler, 2023), and is reported here as a classroom-perceptible pattern across all three disciplinary clusters, with convergent evidence from teacher self-reports and structured observations. Since the study is based on perceptual rather than corpus-level measurement, these results should be interpreted as evidence that instructors perceive and witness homogeneity rather than as direct

language measurement of it. The independently convergent teacher accounts converge on a compound validity threat that hits the construct interpretation node of the TBLA validity chain.

From a psychometric standpoint, if homogenization is occurring at the level teachers report, it would remove the performance variance necessary for a valid interpretation of the structure. When performances converge around a probabilistic mean, the differences observed between students may be attributable to the sophistication of the prompting or access to AI tools, rather than to ESP communicative competence, which would constitute a failure of validity in Messick's (1989) consequential sense (Bachman & Palmer, 2010). The homogenization teachers' report would, from an acquisitional perspective, remove the interlanguage signal that renders learner output diagnostically informative: Corder (1967) established that authentic learner language is necessarily idiosyncratic and variable in ways that reveal developmental trajectory; AI output is statistically probable, not individually variable, and carries no diagnostic information. These failures are made worse by epistemic dependency: students who can't spot genre violations can't leverage them as learning opportunities, and a finding by Ng, Leung, Chu, and Qiao (2021) that critical AI literacy might be among the least-developed AI competences across educational populations offers theoretical support for treating epistemic dependency as a structural, rather than idiosyncratic, risk.

The consistency of findings in all three nodes lends support to a proposition of the TBLT-Assessment-AI Nexus as an initial heuristic model for understanding the disruption of TBLA validity by generative AI at three structurally linked nodes: (1) the task design node, where AI-generated materials may violate the genre-authenticity requirements of ESP TBLA; (2) the assessment validity node, where AI-mediated student performance introduces construct-irrelevant variance violating the fundamental premise of valid assessment; and (3) the construct interpretation node, where linguistic homogenization and epistemic dependency jeopardize the performance variance and interlanguage signal on which valid differentiation and developmental diagnosis depend, according to the teacher accounts on which this study relies. The Nexus is thus provided as a heuristic, not a verified framework, a conceptual map built from the empirical findings of this study, and offered as a scaffold for future research and pedagogical design.

The three nodes are mutually reinforcing: genre-compatible AI materials diminish instructor focus; fluent AI-mediated performances avoid detection; and students never obliged to generate authentic output never gain the genre expertise to evaluate AI critically. All three should be addressed in responses.

At the task design node, the answer is AI-mediated materials evaluation: explicit genre analysis of the incongruence, improper hedging, and disciplinary inaccuracy that AI-generated materials generally produce. into the pedagogy of task design, with instructors taught to detect the register. The answer at the assessment validity node is process-based redesign: multi-stage sequences involving recorded writing, oral defence, and in-class performance that cannot be fully outsourced to AI. This approach is consistent with Norris's (Norris, 2016) idea that TBLA should focus on real-world effects rather than administrative convenience. The response is AI-critical genre literacy instruction at the construct interpretation node. This instruction includes explicit genre analysis tasks, comparative text evaluation, and re-personalisation exercises, which allow students to revise AI-generated drafts to reflect authentic professional discourse and genuine communicative intent.

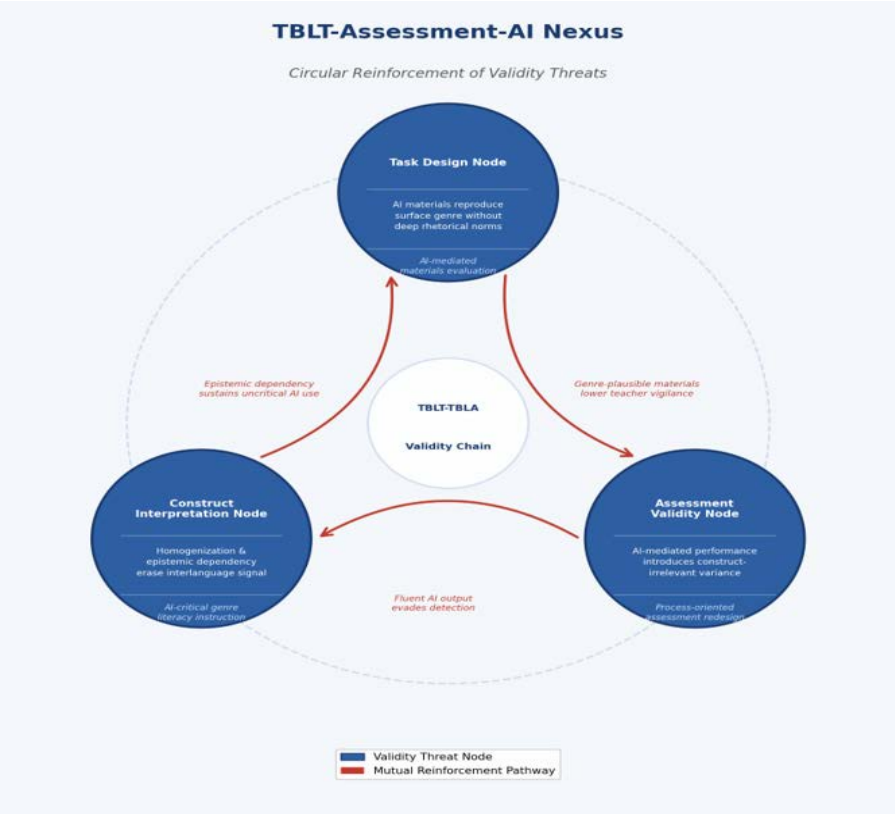


Figure 1. The TBLT-assessment-AI nexus: Three interconnected validity threat nodes

Table 2 clarifies the mapping between the five qualitative content analysis categories and the three Nexus nodes by specifying which categories correspond to each node. Category 1 (Pedagogical Affordances) aligns with the task design node, and Category 2 (Summative Assessment and Validity Concerns) aligns with the assessment validity node; the formative-summative asymmetry observed in Category 2 directly substantiates the theoretical claim associated with the assessment node. Category 4 (Linguistic Homogenization and Epistemic Dependency) is linked to the construct interpretation node, as evidenced by teachers' independent reports of output convergence across disciplinary contexts, which provide the model's strongest empirical support. In contrast, Categories 3 (Teacher Professional Identity) and 5 (Institutional Support Deficits) serve as cross-cutting enabling conditions. These categories operate across all three nodes, determining whether teachers possess the professional resources and institutional support necessary to address Nexus threats. The status of Categories 3 and 5 as enabling conditions is essential: Category 3 suggests that a teacher's professional disposition toward AI—whether experienced as an enhancement or a threat—determines their engagement with node-specific pedagogical responses, such as AI-materials review for task design. Similarly, the lack of AI-critical professional development identified in Category 5 inhibits effective action at both the task design and construct nodes, regardless of an individual teacher's orientation or existing assessment policies. This explicit mapping underscores how each category contributes to or enables responses at specific points along the validity threat continuum.

Table 2. *Correspondence between content analysis categories and TBLT-assessment-AI nexus nodes*

Content Analysis Category	Nexus Node
Category 1: Pedagogical Affordances (n = 38, 95%)	Task Design Node — supports genre-utility claim; AI valued as drafting resource requiring teacher evaluation
Category 2: Summative Assessment & Validity Concerns (n = 35, 87.5%)	Assessment Validity Node — formative/summative asymmetry directly evidences differential validity requirements
Category 3: Teacher Professional Identity (n = 29, 72.5%)	Cross-cutting enabling condition — professional orientation as enabler or threat determines willingness and capacity to enact calibrated pedagogical responses at all three nodes
Category 4: Linguistic Homogenization & Epistemic Dependency (n = 33, 82.5%)	Construct Interpretation Node — independent observations of output convergence constitute strongest empirical support
Category 5: Institutional Support Deficits (n = 40, 100%)	Cross-cutting enabling condition — structural deficit prevents institutional response at any node

The feasibility of the Nexus as an analytical framework rests on three structural properties. First, it is theoretically coherent: each node derives directly from a distinct inferential step in the TBLA validity chain (Bachman & Palmer, 2010; Messick, 1989), ensuring that the model is grounded in established assessment theory rather than ad hoc categorisation. Second, it is empirically complete with respect to the present dataset: the five content analysis categories map exhaustively onto the three nodes (Pedagogical Affordances and Summative Assessment Concerns at the task design and assessment nodes; Homogenization and Epistemic Dependency at the construct node; Professional Identity and Institutional Support as cross-cutting enabling conditions), with no observed phenomenon falling outside the framework. Third, it is non-redundant: each node captures a qualitatively distinct form of validity disruption that the others do not. Genre failure at the task design node does not entail homogenization at the construct node, nor does homogenization at the construct level imply formative-summative asymmetry at the assessment node. Together, these properties establish the Nexus as theoretically viable and empirically grounded, while its designation as ‘preliminary’ and ‘heuristic’ appropriately signals that confirmatory validation across diverse ESP contexts remains necessary.

The most urgent systemic finding is the substantial gap between participants' perceived training needs and institutional policy clarity, a structural deficit that leaves teachers managing the Nexus validity threats without shared conceptual tools or policies, producing fragmented integration practices as the characteristic outcome of professional development neglect (Tondeur et al., 2012).

The qualitative data suggest that formally trained teachers were more confident handling AI-related pedagogical decisions, indicating that professional development of digital competencies is necessary and demonstrably effective. Critically, this confidence extended specifically to summative assessment contexts: teachers who had engaged with AI tools professionally demonstrated a sharper capacity to identify AI-inflected student output, to recognise the characteristic flatness of probabilistic register, and to cross-examine fluent performances that lacked the individual reasoning and field specificity of authentic learner production. AI literacy, in other words, functions as a form of assessment literacy: the teacher who understands how large language models generate text is better positioned to detect when a student has delegated rather than produced, and to make valid summative judgements accordingly. Liu's (2025) AI literacy framework specifies the competency content required; this study adds the pedagogical sphere: task-based assessment design in ESP contexts, where genre knowledge and validity awareness

must be more exacting than in general-language teaching. Case et al.(2025) formulation applies directly: the goal is critical pedagogical judgment, not mere tool competence.**Conclusion**

This study offers empirical evidence and conceptual understanding of the impact of generative AI on ESP pedagogy, task-based language teaching, and assessment validity. University ESP teachers consistently differentiated GenAI's pedagogical value in task design and formative feedback from its potential for summative assessment, as AI-assisted production has been shown to undermine construct validity. They also noted a growing linguistic homogenization and epistemic dependence among students, indicating that reliance on generative AI may replace the actual participation and productive struggle necessary to develop disciplinary communicative competence. This TBLT-Assessment-AI Nexus combines these findings, connecting task design, assessment validity, and construct interpretation, and suggests the need for systemic AI policies and not a teacher-led response. While more longitudinal, corpus-based, and cross-national research is needed, our findings suggest that education systems should offer students the opportunity to use generative AI as a framework for learning rather than as an alternative to genuine language engagement and to provide an authentic, proficient language.

Funding: This research is not funded.

Acknowledgements: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Authenticity: This manuscript is an original work.

Artificial Intelligence Statement: The authors confirm that AI-assisted tools, including Grammarly and Smodin, were used solely for language checking, editing, and formatting assistance. Following the use of these tools, the authors reviewed and edited the manuscript as necessary and take full responsibility for the integrity and accuracy of the work as published.

About the Authors:

D.Sc., Prof. Nataliia Avsheniuk is a distinguished scholar in education, Professor and Head of the Foreign Systems of Pedagogical and Adult Education Department at the Ivan Ziaziun Institute of Pedagogical and Adult Education of the National Academy of Educational Sciences (NAES) of Ukraine. Her research interests focus on comparative education, adult education, quality assurance in higher education, lifelong learning policies, and pedagogical innovations in global educational contexts. ORCID ID: https://orcid.org/0000-0001-6246-4132

Nataliia Seminikhyna, PhD in Education, Associate Professor. Her research interests include English for Specific Purposes (ESP) teaching methods, language pedagogy, comparative pedagogy, teacher education, teacher professional development, and the integration of artificial intelligence into language teaching and assessment. ORCID ID: https://orcid.org/0000-0001-6246-4132

Olena Lutsenko, PhD in Education, Associate Professor. Her research interests include English for Specific Purposes (ESP) teaching methods, language pedagogy, and teacher education. She has

contributed to various publications in the field of education and ESP teaching, focusing on innovative approaches to foreign language instruction and professional development in higher education. ORCID ID: 0000-0002-6864-908X

Tetiana Svyrydiuk, PhD in Linguistics, Associate Professor. Her research interests involve cognitive linguistics, applied linguistics, cross-cultural communication, and ESP teaching. She has extensive experience in foreign language instruction, with a focus on developing communicative competence and cross-cultural awareness. ORCID ID: https://orcid.org/0000-0001-9482-8228

Larysa Ruban, PhD in Education, Associate Professor. Her research interests involve English for Specific Purposes (ESP), teacher education, and teacher professional development. She has extensive experience in ESP teaching and has contributed to various publications in the field of education and foreign language instruction. ORCID ID: https://orcid.org/0000-0001-5539-931X

References

Avsheniuk, N., Seminikhyna, N., Ruban, L., & Sviatiuk, Y. (2025). Exploring overreliance on AI tools in English for Specific Purposes courses: Challenges and implications for learning and academic integrity. *Arab World English Journal (AWEJ) Special Issue on Artificial Intelligence*, 3–20. https://dx.doi.org/10.24093/awej/AI.1

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Bhandari, L. P., Dahal, N., Awasthi, J. R., & Dhungana, S. (2025). Technology-mediated task-based language teaching: A systematic review. *Cogent Education, 12*(1), Article 2560051. https://doi.org/10.1080/2331186X.2025.2560051

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Longman.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT-generated AI writing and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), Article 100068. https://doi.org/10.1016/j.rmal.2023.100068

Case, R., Liu, L., & Mintz, J. (2025). Integrating AI technology into language teacher education: Challenges, potentials, and assumptions. *Computers in the Schools, 42*(2), 93–99. https://doi.org/10.1080/07380569.2025.2458950

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education. *Educational Technology & Society, 25*(1), 28–47. https://www.jstor.org/stable/48647028

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*(4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ertmer, P. A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational Technology Research and Development, 53*(4), 25–39. https://doi.org/10.1007/BF02504683

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Hyland, K. (2019). Genre and second language academic writing. *Language Teaching, 52*(2), 149–163. https://doi.org/10.1017/S0261444818000429

Li, Y., Shadiev, R., & Chiu, T. K. F. (2025). Applying generative artificial intelligence to task-based language teaching and learning: A systematic review and meta-analysis. *TechTrends,* 70*(1).* https://doi.org/10.1007/s11528-025-01140-7

Liu, W. (2025). Language teacher AI literacy: Insights from collaborations with ChatGPT. *Journal of China Computer-Assisted Language Learning*. Advance online publication. https://doi.org/10.1515/jccall-2024-0030

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3313831.3376727

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley-Blackwell.

Meniado, J. C. (2023). The impact of ChatGPT on English language teaching, learning, and assessment: A rapid review of literature. *Arab World English Journal, 14*(4), 3–18. https://dx.doi.org/10.24093/awej/vol14no4.1

Messick, S. (1989). In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). SAGE.

Ministry of Education and Science of Ukraine [MES Ukraine]. (2024). *Recommendations for the use of artificial intelligence in the educational process of general secondary education institutions*. https://mon.gov.ua/ua/news/rekomendaciyi-shodo-vikoristannya-shtuchnogo-intelektu-v-osviti

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualising AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578–594). Wiley-Blackwell.

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230–244. https://doi.org/10.1017/S0267190516000027

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Palgrave Macmillan.

Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tondeur, J., van Braak, J., Sang, G., Voogt, J., Fisser, P., & Ottenbreit-Leftwich, A. (2012). Preparing pre-service teachers to integrate technology in education: a synthesis of

qualitative evidence. *COMPUTERS & EDUCATION*, *59*(1), 134–144. https://doi.org/10.1016/j.compedu.2011.10.009

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Valerio, M. T. B. (2025). Enhancing Argumentative Essay Skills: Evaluating the Efficacy of AI-Powered Tools Versus the Traditional Approach Among Learners. *Arab World English Journal (AWEJ) Special Issue on Artificial Intelligence*: 270- 289. https://dx.doi.org/10.24093/awej/AI.15