

The Effectiveness of NotebookLM Podcasts in ESP Classes

Daniela Sršníková¹
Ivana Pobočíková²

¹ University of Žilina in Žilina, Faculty of Mechanical Engineering, Univerzitná 8215/1, 010 26 Žilina, daniela.srsnikova@fstroj.uniza.sk

² University of Žilina in Žilina, Faculty of Mechanical Engineering Univerzitná 8215/1, 010 26 Žilina, ivana.pobocikova@fstroj.uniza.sk

Grant: 025ŽU-4/2024

Name of the Grant: Implementation of new didactic tools to increase the quality of mathematics teaching in engineering master's degree programs at technical universities

Category: AM – Pedagogy and education

© GRANT Journal, MAGNANIMITAS Assn.

Abstract The study evaluates the contribution of AI-generated podcasts from Google NotebookLM to English for Specific Purposes (ESP) instruction. A six-week quasi-experiment examined listening comprehension and technical vocabulary acquisition among 20 engineering students at the Faculty of Mechanical Engineering at the University of Žilina in Žilina. Technical texts were converted to podcast format using the Audio Overview feature. Student progress was assessed through three tools of increasing cognitive demand: a listening comprehension test, an inferential reasoning task, and a written summary requiring productive use of terminology. The findings suggest that AI-generated podcasts can effectively support listening development and terminology acquisition, provided that the pedagogical approach accounts for differences between proficiency levels.

Key words NotebookLM, ESP listening comprehension, terminology acquisition, AI-generated podcast, cognitive load

1. INTRODUCTION

Students at the Faculty of Mechanical Engineering, University of Žilina, face several challenges that present a significant pedagogical issue for teachers of English for Specific Purposes (ESP). One identified challenge is the insufficient availability of authentic audio materials targeting the specific domains of mechanical engineering. Commercially available audio materials are typically too general and do not contain the technical terminology and discipline-specific discourse required in ESP instruction. Audiovisual materials documenting technological processes, by contrast, offer valuable visual context but frequently contain only limited spoken specialist discourse. ESP teachers, therefore, find themselves in a position where they must either work with materials that do not match the professional profile of their students or invest considerable time in developing their own resources. As a result, students have limited opportunities for the systematic development of listening skills and the acquisition of terminology through audio materials. The insufficient availability of such materials consequently restricts the development of their professional communicative competence in English.

This identified gap prompted an investigation into the potential of the Google NotebookLM platform as a tool for creating audio materials. The platform's Audio Overview feature enables

the transformation of technical texts into podcast format. Unlike conventional generative tools, the outputs generated by NotebookLM are grounded in the source documents, which, according to available studies, reduces the risk of hallucinations and contributes to the preservation of domain-specific terminological register (Reyna, 2025; Tufino, 2025). The platform, therefore, represents a promising means of addressing the identified gap.

Before integrating the platform into instruction in the winter semester of 2026, we first considered it essential to evaluate its effectiveness to ensure that it would meaningfully support students' learning. The current research literature does not provide convincing evidence that AI-generated podcasts develop language skills. There are additional grounds for caution. Albrecht-Crane (2025) warns that overly fluent and frictionless content may reduce students' cognitive effort. When content does not present an adequate challenge, students may process it in a predominantly passive manner. Passive reception of information alone, however, is not sufficient for effective learning. From the perspective of cognitive psychology, managing an appropriate level of cognitive load is a necessary precondition for the durable acquisition of linguistic structures. In other words, the high fluency of AI-generated content may create an illusion of learning, leading students to believe they have understood the material even though they have not engaged in the deeper cognitive processing necessary for meaningful learning.

This tension between ease of engagement and the conditions required for learning became the central question of our study. Over six weeks, we examined whether the regular integration of AI-generated podcasts into ESP instruction improved listening skills and terminology acquisition among B1- and B2-level students. Student progress was assessed through three tasks of increasing cognitive demand. The first verified basic comprehension of the audio text. The second targeted the ability to draw inferences not explicitly stated in the podcast. The third task assessed the ability to use terminology productively in written output. The combination of these three metrics enables distinguishing between situations in which AI-generated podcasts genuinely support learning and those in which they create the appearance of learning.

2. COGNITIVE CHALLENGES OF AI

Whether artificial intelligence supports learning depends on its capacity to generate contextualised and adaptive interaction, rather than on the mechanical transmission of knowledge. Therefore, the value of modern educational technology lies not in providing ready-made information, but in its capacity to adapt dynamically to learners' needs in real time. This premise is supported by Kestin et al. (2024), who found that personalised learning through AI tutors had significantly better post-test results than traditional forms. Although that research was conducted in the exact sciences, its methodological conclusions carry broader didactic relevance. They suggest that when technology functions as a partner in the learning process, it can substantially increase instructional effectiveness. This capacity for adaptive interaction provides the foundation for contemporary foreign language teaching methodologies designed to develop complex communicative competencies.

This opening argument leads directly to the question of technological reliability and accuracy, which is particularly important in ESP instruction. The integration of platforms based on retrieval-augmented generation (RAG) architecture represents a significant step towards ensuring terminological consistency and subject-matter precision. Unlike generative models that draw on open data sources, RAG-based systems restrict output generation exclusively to documents explicitly uploaded by the user. Reyna (2025) situates Google NotebookLM within the paradigm of adaptive learning and emphasises that its RAG architecture anchors all outputs to user-provided sources, substantially reducing the risk of hallucinations. In the context of ESP instruction, this feature is particularly significant, as factual and terminological inaccuracies represent the main barriers to trust in AI tools. Addressing these risks creates a more reliable environment in which students can confidently develop terminology and professional competencies.

Despite these advantages, the formal fluency of AI outputs carries a risk of cognitive offloading that may paradoxically reduce critical thinking and lower language acquisition. When technology assumes too large a share of the cognitive load, it can weaken the active processing of information that sustained learning requires. Polverini and Gregorcic (2024) highlight the limitations of AI as an autonomous problem-solving tool, while Albrecht-Crane (2025) and Gerlich (2025) further point out that the immediate availability of answers may reduce students' cognitive engagement and limit deeper learning processes. In language learning, this issue relates to Krashen's (1985) input hypothesis ($i+1$), which proposes that meaningful progress occurs when learners are exposed to language input that slightly exceeds their current level of competence. This tension between technological fluency and the requirement for cognitive challenge forms the core of our study. For this reason, relying solely on assessments of basic text comprehension is insufficient, given that such measures may be confounded by passive content reception. Therefore, we decided to examine the depth of inferential processing and students' ability to apply terminology productively in new contexts. Only the combination of these three metrics enables us to determine whether AI-generated content within the NotebookLM environment supports the development of listening skills and terminology.

3. METHODOLOGY

3.1 Research Design and Hypotheses

The primary methodological aim of the study was to verify the effectiveness of AI-generated podcasts for developing listening skills and acquiring mechanical engineering terminology

in a university context. To capture the fine-grained temporal dynamics of student progress, all three assessed metrics were monitored longitudinally over six weeks. Effectiveness was evaluated based on the systematic increase in students' scores across the assessment period. For each group and each week, basic descriptive statistics were calculated, specifically the mean and standard deviation, both for the sample as a whole and separately for the B1 and B2 proficiency levels. This step enabled the formulation of two research hypotheses. The first hypothesis tested whether exposure to the podcasts produced a statistically significant monotonically increasing trend in weekly scores, examined separately for the B1 and B2 groups. The second hypothesis examined whether both groups followed the same developmental trajectory over time or whether their learning curves diverged to a statistically significant extent.

3.2 Statistical Methods

To test the first hypothesis, Page's trend test was selected in preference to the more commonly used Friedman test. This choice was methodologically justified by the greater statistical power of Page's test for verifying pre-specified ordinal hypotheses that assume a clearly defined ascending trend from W1 to W6 (Hendl, 2006; Rublík, 2011). Testing was conducted independently for each metric and for each proficiency group at a significance level of $\alpha = 0.05$. As no single standardised effect size index exists for Page's test, practical significance was quantified using the standardised z-score derived from Page's L statistic (Peng and Chen, 2015). The generalised Spearman rank correlation coefficient ρ was calculated, and results were interpreted according to Cohen's scale (Cohen, 1988; Field, 2026).

The second hypothesis was tested through profile analysis (Anděl, 1998; Lehmacher, 1978). Profile analysis enabled a comparison of the score development trajectories of the two independent groups by examining their parallelism, overall level, and changes over time. When the analysis revealed differences between the groups' score trajectories, it was further used to determine the specific weeks in which statistically significant divergences occurred. All tests were conducted at a significance level of $\alpha = 0.05$. The profile analysis allows the six-week process to be treated as a single integrated dynamic system. This comprehensive statistical approach therefore provides methodological evidence for determining whether AI-generated podcasts reduce differences among students or whether the benefits of their implementation remain dependent on students' initial levels of language proficiency.

3.3 Participants and Intervention Procedure

The research sample consisted of 20 students enrolled in the degree programmes Mechanical Engineering Technologies and Computer-Aided Design and Simulation, taking the compulsory course Technical English for Engineers 2 at the University of Žilina. Participants' initial level of language proficiency was assessed using the standardised Oxford Quick Placement Test (OQPT), which provided the data necessary to divide the sample analytically into two equal groups reflecting distinct proficiency levels: ten students at B1 level with a mean score of 32.4 (SD = 2.1), and ten students at B2 level with a mean score of 44.7 (SD = 3.3).

Each week constituted one measurement cycle. Teachers uploaded a weekly text compiled from authentic English-language engineering articles to NotebookLM and used it to generate a two-host conversational podcast with an average duration of 9.4 minutes. Each podcast was checked for factual accuracy before deployment,

following the quality-control protocol recommended by Tufino (2025). Students listened to the podcast twice. The first listening focused on comprehension, the second on noting terminology. Student progress was quantified using three assessment metrics with progressively increasing cognitive demands, enabling the tracking of development from accurate recognition of linguistic forms and inferential reasoning to the productive use of terminology.

3.4 Metrics

The first metric was a combined listening comprehension test (M1), designed to assess two different levels of receptive sub-skills. The metric comprised two task types, a cloze test and a multiple-choice test, to eliminate the risk of distortion from a one-sided methodology. The cloze task contained ten gaps selected from the podcast transcript. For example, in the unit on ultrasonic machining, one item stated: *"In ultrasonic machining, the abrasive ____ is circulated between the vibrating tool and the workpiece,"* requiring students to identify the precise technical term *slurry*. The multiple-choice test assessed global comprehension through questions such as *"What is the main advantage of ultrasonic machining over conventional machining according to the podcast?"*, where the selection of the correct option (c), *the ability to machine hard and brittle materials without significant thermal damage*, depended on following an argument developed across multiple speaking turns. The combination of these tasks made it possible to capture not only the immediate decoding of a specific linguistic form, but also the student's capacity to reconstruct broader informational relationships within discourse. Metric M1 thus provided a reliable quantitative basis for evaluating listening comprehension and the basic retention of explicitly presented information.

The second metric was a five-item inferential reasoning task (M2), designed to elicit a qualitatively distinct and cognitively deeper level of processing. The defining constraint was that the correct answer to each question was never stated directly in the podcast. Each item required the student to independently derive a logical conclusion that was implicitly entailed by the content but never verbally expressed. One item from the ultrasonic machining unit illustrates this: *"The podcast states that ultrasonic machining removes material through abrasive action rather than cutting and produces minimal heat... What can you infer about its suitability for machining heat-sensitive materials compared to conventional machining?"* Since the podcast does not directly address the suitability for heat-sensitive materials and only describes the material removal mechanism, arriving at the correct solution requires students to integrate multiple pieces of information presented across different sections of the podcast. Whereas M1 primarily assessed auditory memory, M2 examined deeper comprehension by determining whether students could independently process a received concept and transform it into new logical relationships. Therefore, the introduction of the inferential metric M2 marked a substantive cognitive boundary between passive text decoding and active intellectual processing.

The third and most demanding metric was a written summary of the podcast content in the student's own words (M3), evaluated according to the frequency and appropriateness of productive use of target terminology. Unlike the preceding tasks, this exercise was completed without access to notes or written support to maximise the demands placed on the student's ability to convert passively retained lexical units into active vocabulary. In the unit on ultrasonic machining, the monitored terminological register included the terms *abrasive slurry*, *vibrating tool*, *material removal rate*, *brittle materials*, and *surface finish*. The difference in production quality

was pronounced. A B2-level student integrated the terminology into a structurally complex response: *"In ultrasonic machining, a vibrating tool transmits high-frequency oscillations to the abrasive slurry, so we can machine hard and brittle materials with a perfect surface finish"*, achieving a high type-token ratio (TTR). A lexically impoverished response with a low TTR, by contrast, consisted only of the mechanical repetition of basic phrases: *"The tool vibrates. The machine removes material."* Within the overall research design, metric M3 served as the final productive indicator of terminological acquisition.

4. DISCUSSION

Initial analysis of the obtained data indicated a clear performance differentiation between the two proficiency levels, with B2 students consistently achieving higher mean scores across all three metrics. As shown in Table 1, for basic listening comprehension (M1), the B2 group achieved a mean score of 72.89 (SD = 4.58), while the B1 group value was 63.69 (SD = 4.58). A similar but more significant gap was for inferential accuracy (M2), where B2 students achieved a mean of 54.51 (SD = 9.21), while the B1 group stagnated at 39.49 (SD = 5.33). The most substantial difference was observed in productive lexical density (M3). The B2 group absorbed and applied terminology at a mean level of 0.48 (SD = 0.38), whereas the B1 group showed near-zero initial and mid-term production, with an overall mean of only 0.03 (SD = 0.38). These data show that task cognitive demand amplifies the gap between the groups. Whereas for passive decoding (M1) the difference in means is approximately 9 points, for complex deductive reasoning (M2) and active vocabulary use (M3) the asymmetry becomes the dominant characteristic of the distribution.

Table 1: Descriptive statistics for individual metrics in the B1 and B2

Metric	Mean B1	SD B1	Mean B2	SD B2	Mean B1+B2	SD B1+B2
M1	63,69	4,58	72,89	4,58	68,29	7,04
M2	39,49	5,33	54,51	9,21	46,99	10,63
M3	0,03	0,38	0,48	0,38	0,26	0,36

A detailed examination of results across the time horizon from week one to week six (W1 to W6) revealed a consistent upward trend in mean scores for both groups, though with differing internal dynamics. This increase is captured in Table 2 and visually verified by the figures (Figures 1 to 3), which confirm a steady progression towards higher performance. Analysis of basic comprehension (M1; Figure 1) indicates that both groups experienced a significant increase between the initial two weeks (W1–W2) and the final two weeks (W5–W6), with the B1 group reaching a final score of 71.69 and the B2 group 76.98. In inferential reasoning (M2; Figure 2), the B1 group demonstrated progress mainly between the initial phase and the final week, achieving a score of 43.64, whereas the B2 group showed a stronger and more consistent upward trajectory, reaching 67.51 in W6. The most critical pattern was observed for lexical production (M3; Figure 3). B1 students demonstrated no productive lexical output from W1 to W4; the first evidence of actual term acquisition appeared only in W5 (0.01), followed by a substantial increase in W6 (0.19). The B2 group, by contrast, accelerated productive term use continuously from W1 (0.06) through to a final value of 0.94 in W6. This weekly distribution reveals a key analytical finding: the growth trend is not internally homogeneous. The stagnation of the B1 group means for metric M3

during the first four weeks provides both visual and mathematical evidence of a latent learning phase, in which the absence of visible performance masks an internal accumulation of knowledge that was fully released only in the final block. Analysis of the weekly profiles clearly shows that AI-generated podcasts stimulate progress in all students. The trajectory of the transformation from passive listening to active written production is non-linear and dependent on the student's initial level of proficiency.

Table 2: Descriptive statistics for individual metrics by week

	M1		M2		M3	
	Mean B1 (SD)	Mean B2 (SD)	Mean B1 (SD)	Mean B2 (SD)	Mean B1 (SD)	Mean B2 (SD)
W1	56,36 (4,35)	67,70 (3,24)	37,46 (3,37)	41,97 (2,97)	0 (0)	0,06 (0,08)
W2	59,40 (3,51)	69,49 (2,77)	34,54 (3,67)	47,81 (3,99)	0 (0)	0,31 (0,25)
W3	62,83 (2,90)	73,22 (4,49)	38,04 (5,17)	52,99 (4,13)	0 (0)	0,28 (0,24)
W4	63,36 (3,18)	73,48 (3,97)	40,30 (2,83)	56,28 (3,90)	0 (0)	0,45 (0,21)
W5	68,52 (1,80)	76,47 (2,31)	42,94 (5,57)	60,47 (4,97)	0,01 (0,03)	0,84 (0,26)
W6	71,69 (2,92)	76,98 (2,10)	43,64 (4,19)	67,51 (3,84)	0,19 (0,19)	0,94 (0,19)

4.1 Metric 1: Listening Comprehension

Longitudinal analysis from W1 to W6 revealed a consistent upward trend in both groups. For basic comprehension (M1), the B1 group ended the experiment at 71.69 and the B2 group at 76.98. Page's trend test confirmed a highly significant upward direction over time for both groups. The B1 group displayed a pronounced ordinal trend: $L = 888$; $p < 0.001$; $ES = 6.18$; $\rho = 0.87$, indicating a large effect size. The B2 group achieved an analogously significant, though more moderate, upward trend: $L = 855$; $p < 0.001$; $ES = 4.85$; $\rho = 0.69$. Profile analysis confirmed that the growth curves of both groups followed the same trajectory over time. Both groups thus progressed at a comparable rate, regardless of entry-level proficiency. The larger effect in the B1 group suggests that the audio input held higher relative value for these students and functioned as a more powerful accelerator of progress. The results for metric M1 confirm that AI-generated podcasts function as a reliable tool for improving basic comprehension across the full sample.

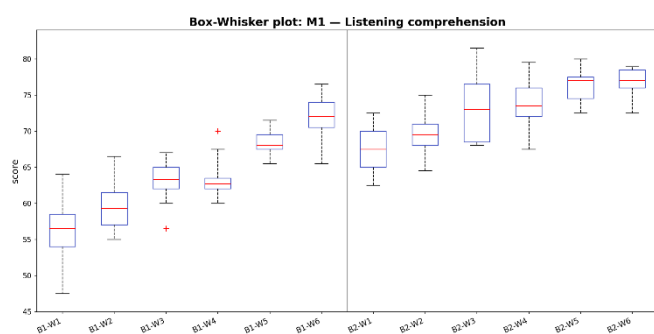


Figure 1. Weekly M1 score values for the B1 and B2 groups.

4.2 Metric 2: Inferential Accuracy

Analysis of the ability to draw logical conclusions from text (M2) revealed a different dynamic. Although Page's test confirmed a significant upward trend in both groups, the B2 group showed a considerably steeper rate of progress than the B1 group. The B2 group achieved maximum effect strength: $L = 896$; $p < 0.001$; $ES = 6.51$; $\rho = 0.92$. The B1 group demonstrated a substantially weaker, though still significant, progression: $L = 836$; $p < 0.001$; $ES = 4.08$; $\rho = 0.58$, which on Cohen's scale corresponds to a medium effect size. Profile analysis confirmed that the developmental trajectories of the two groups differed significantly, with the key divergence point identified as early as W1. The explanation, in our view, is straightforward. Logical inference places greater cognitive demands on students than the first task type. The continuous speech stream placed demands on the B1 group that exceeded their capacity for deep processing. In contrast, the B2 group possessed sufficient linguistic proficiency to draw implicit inferences from the beginning and therefore benefited more from this task. The implication for pedagogical practice is unambiguous. The effectiveness of AI-generated podcasts for cognitively demanding tasks depends largely on students' initial language proficiency, which affects both comprehension depth and active content processing. B1-level students should therefore receive additional methodological support, for example, pre-listening activities to help them bridge the linguistic gap.

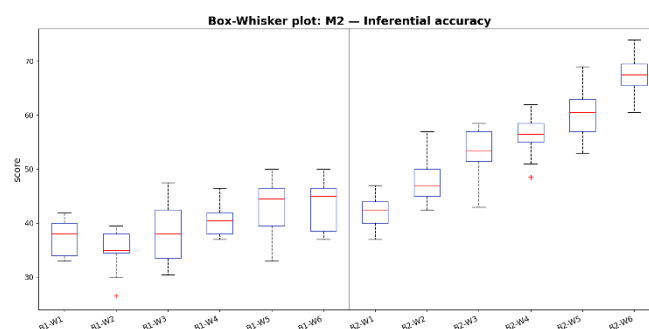


Figure 2. Weekly M2 score values for the B1 and B2 groups.

4.3 Metric 3: Lexical Density

The greatest asymmetries in learning trajectories emerged for productive use of terminology (M3). The B2 group showed linear growth from W1 (0.06) to W6 (0.94). Page's test confirmed a highly significant trend with a large effect size: $L = 876.5$; $p < 0.001$; $ES = 5.72$; $\rho = 0.81$. The B1 group followed a fundamentally different course. From W1 to W4, production remained at zero. The first evidence of active use of terminology did not appear until W5 (0.01), followed by a significant increase in W6 (0.19). Page's test for the B1 group revealed only small practical significance: $L = 799.5$; $p = 0.0092$; $ES = 2.61$; $\rho = 0.37$. Profile analysis confirmed that the developmental trajectories diverged significantly at both W1 and W5, after which W6 saw a sharp convergence. We describe this behaviour as cognitive incubation. New terminology remained stored in passive memory for an extended period and was transferred into written production only after six weeks of listening. The B2 group consistently transformed auditory input into productive language output throughout the study. The B1 group exhibited a prolonged period of apparent stagnation before showing a significant increase in performance. This asymmetry has important implications for interpreting the results. Had the intervention lasted only four weeks, the B1 group's M3 performance might have been

misinterpreted as regression, when it reflected a transitional phase of intensive but still latent language acquisition.

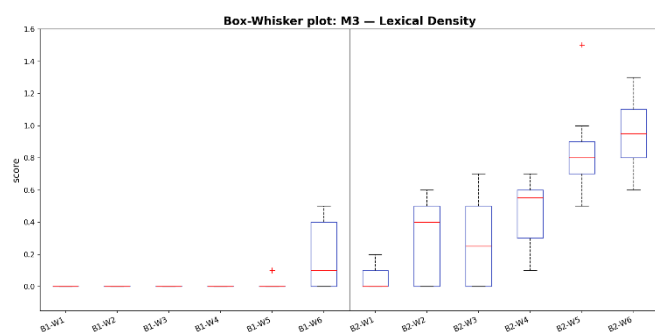


Figure 3. Weekly M3 score values for the B1 and B2 groups.

5. STUDENTS' PERSPECTIVES

5.1 Data Collection

After the six-week intervention, a qualitative focus group interview was conducted with all twenty participants. The discussion was held in Slovak, lasted approximately 60 minutes, and was audio-recorded for subsequent textual transcription. The role of moderator was taken by the teacher, who guided the discussion using open questions organised around three main areas: students' personal experience of listening to the podcasts, their perceptions of the specific assessment tasks, and their overall evaluation of the intervention. Through these areas, we examined more closely how students perceived the difficulty of the tasks, what they found supportive, and where they encountered obstacles. The aim of this procedure was not to quantify students' responses, but to gain a comprehensive understanding of how they experienced each task.

5.2 Data Processing

Student responses were analysed using the established methodology of thematic analysis as set out by Braun and Clarke (2006). In the initial phase, the audio recording was transcribed, and each statement was assigned a descriptive code capturing its core meaning. In the second step, these codes were systematically organized into broader thematic categories according to the assessment task they addressed and the type of student experience they reflected. This process identified three themes: perceived task manageability, cognitive load during content processing, and the relationship between passive comprehension and active production. Representative student statements are presented in their original form, each accompanied by a brief interpretive commentary. This final step situates students' qualitative accounts within the broader context of the quantitative findings.

5.3 Metric 1: Listening Comprehension

Most participants described the M1 task as requiring little additional effort, largely because they were already familiar with this type of task format. This pattern is confirmed by typical student responses, in which the word-completion mechanism was perceived as automatic: *"I am used to this type of task. We did it at primary school and basically almost every time we did listening..."* Another student added: *"This task is very easy for me, you just listen and write the missing word... You don't even need to understand*

everything; you just need to hear the order of the words." Quantitatively, performance on this task differed little between the B1 and B2 groups, and both groups followed similar growth trajectories. However, the qualitative findings revealed an important difference in how students experienced the task. Several B1-level students reported initial difficulty keeping pace with the speed of delivery: *"We are really trained for this type of task, but the speed seemed higher to me... You didn't stop the recording after each gap, so I couldn't quite manage to write the word... and listen to what was coming next at the same time."* Detailed analysis of these responses indicates that the obstacle was not the task type itself, but the cognitive load associated with writing simultaneously while listening continuously without pauses. This detail ultimately makes clear that although task M1 did not differentiate the proficiency groups statistically, the experience of the listening process was not the same for both groups of students.

5.4 Metric 2: Inferential Accuracy

With task M2, the responses and strategies of B1 and B2 students began to diverge markedly. While students in the B2 group described the task as intellectually stimulating and as requiring an active change of approach, the B1 group encountered its cognitive limits. Authentic responses from both groups clearly illustrate this contrast. B2 respondents reflected on the need for deeper textual comprehension and deductive reasoning: *"I enjoyed this task, but it also frustrated me... I was really pleased when I got full points, because this isn't about parroting, you must understand,"* while several of them reported that during the second listening they began actively searching for information: *"I didn't realise I had to deduce something different from the individual pieces of information... But once I understood what you wanted... I already knew while listening, right, so this is probably going to be there."* Among B1 students, by contrast, a persistent and frustrating sense prevailed that the text did not contain the answer: *"I didn't like this task... I kept feeling that the information wasn't there, even though I felt I understood what was being said."*

This perception of missing information among the B1 group was not attributable to unclear task instructions. Instead, it reflected a depletion of cognitive resources during processing. At the theoretical level, this mechanism is precisely predicted by Krashen's (1985) input hypothesis (i+1), according to which language acquisition occurs when input only slightly exceeds current competence. When it exceeds that competence substantially, the capacity for deeper content processing is lost. In terms of cognitive load theory (Sweller, 1988), working memory has a limited capacity, and in B1 students, the lexical and syntactic decoding of spoken discourse consumed so large a share of that capacity that inferential operations, the drawing of conclusions not directly stated in the text, became cognitively unavailable to them.

The outcome of incorrect or absent responses among the B1 group was therefore not caused by a failure to understand the nature of the task, but by a limit on working memory capacity. This distinction has direct implications for pedagogical practice. The appropriate response is not to simplify the inferential tasks themselves, but to reduce the cognitive load associated with language decoding. This can be achieved by incorporating pre-listening activities that activate the necessary cognitive schemas and free up students' mental capacity for more demanding reasoning operations.

5.5 Metric 3: Lexical Density

Students' responses to task M3 offered the deepest insight into the learning process, revealing a clear gap between immediate productive capacity and the process of cognitive incubation. B2-level students described this synthesis task as a natural and liberating extension of listening: *"This suited me most, because I could show that I really understand... I wasn't constrained in any way... I use the technical words in my sentences"*, for the B1 group, M3 was primarily a source of pronounced frustration during the first four weeks (W1 to W4). This initial phase was characterised by zero or minimal production, which some students attributed to the language barrier or to a pragmatic view of their future profession: *"I didn't like this exercise... I'm fine with understanding that on the shop floor you don't need to communicate... I know the terms, and I can explain the process, but in Slovak."*

Gradually, however, a significant shift occurred in a portion of the B1 students, as they came to understand that the goal was not grammatical complexity but terminological precision: *"...gradually it started working, when I realised that I didn't need to use those complex tenses, it's enough to capture the essence, but with the right terms. Then it made sense why I should use sintering instead of baking."* These statements are the most precise verbal expression of the phenomenon of cognitive incubation. The passive memory trace was genuinely present in the students' minds, but active production required time. The statistical inflection point that analysis identified at W5 and W6 was described by students themselves at the end of the semester as a moment of surprise at their own progress: *"...at the last lesson I got the most points on the final exercise, and basically I just combined what we had learned in technology classes with the vocabulary you wanted from us... that task seemed the best to me, because I could write what I wanted."*

This observation has a central pedagogical implication for the design of ESP instruction. It demonstrates that zero productivity in the early phases of the intervention is not a failure of the method, but a necessary latent period during which new linguistic material is being internalised in memory. If a teacher does not recognise this specific dynamic process, the initial silence may be misread as failure, prompting the premature introduction of excessive scaffolding. Interrupting this process prematurely cuts short the necessary absorption period, preventing a deeper and more lasting understanding of the new technical vocabulary.

6. CONCLUSION

The study produced three significant and interconnected empirical findings regarding the integration of AI tools into ESP instruction. The results show that although NotebookLM AI-generated podcasts demonstrably improve students' listening comprehension and terminology acquisition, the nature and pace of this progress vary considerably according to their initial language proficiency. Quantitative analysis confirmed statistically significant progress across all three metrics (M1, M2, M3); however, the B1 and B2 groups responded to identical audio input with different dynamics, and the divergence intensified as the cognitive demand of the tasks increased. This variation was most significant in the written summary task (M3), where the B1 group experienced an extended latent phase followed by a marked improvement in performance during the final week of the intervention. This pattern of cognitive incubation demonstrates that the absence of immediate visible performance gains does not necessarily indicate an absence of learning. For educational research, this carries a consequential warning against the risk of premature evaluation of AI tool

effectiveness, which may lead to the unwarranted rejection of innovative pedagogical methods.

The quantitative growth trajectories are further explained by qualitative feedback, which reveals how students themselves perceived changes in their knowledge development. Subjective reflections collected at the end of the intervention confirm that the three assessment tasks elicited different forms of engagement, with responses varying markedly not only between the B1 and B2 proficiency levels but also within each group. While most students perceived the listening comprehension (M1) as manageable or even routine, requiring only memory, the inferential reasoning task (M2) and written summary (M3) became explicit points of cognitive conflict. B2-level students described inferential analysis (M2) as intellectually stimulating, whereas B1-level students experienced the same task as cognitively overloading. Their ability to draw inferences was constrained by the substantial linguistic demands to decode the language itself. Therefore, the perceived fluency of M1 constitutes a pedagogical risk, consistent with Albrecht-Crane's (2025) argument that a task perceived as too easy may not generate the depth of information processing required for durable knowledge consolidation.

From the perspective of the acquisition process, the most notable feature was the connection between statistical stagnation and students' emotional experience of the written summary (M3). Data analysis showed that this productive task divided the two groups most sharply. Furthermore, for B1-level students, it was a primary source of frustration in the early weeks. Students in this group repeatedly reported that although they understood the podcast content, they were unable to transform it into their written production. A turning point occurred in weeks five and six, when respondents began spontaneously describing the moment when terms they had previously only recognised started flowing naturally into their written output. This qualitative transformation provides direct evidence of a latent learning phase. Furthermore, it confirms that the emotions associated with initial task difficulty represent a natural part of the cognitive restructuring that precedes the development of higher levels of language competence.

The findings carry direct and applicable implications for curriculum design and pedagogical practice in ESP. The research demonstrates that while AI-generated podcasts function as a reliable tool for developing basic comprehension independently of proficiency level, a differentiated pedagogical approach is necessary for cognitively more demanding tasks. In practice, this means that for B1-level students, teachers must incorporate targeted pre-listening tasks designed to reduce cognitive load and activate relevant lexical schemas. For B2-level students, podcasts can serve as a primary source of acquisition input without additional scaffolding if assessment tasks require the productive use of terminology rather than its passive recognition. A correct understanding of the latent period also prevents premature teacher intervention, giving students sufficient time for deeper and more durable knowledge consolidation.

Despite these contributions, the conclusions presented here should be interpreted in the context of several methodological limitations. The research design was constrained by the absence of a control group, a relatively small sample size ($n = 20$), and a six-week intervention period without a delayed post-test. These limitations prevent the formulation of unequivocal causal conclusions and restrict the assessment of long-term retention. Future research should address these issues through replication studies including control groups, larger samples, and longitudinal measurements with extended follow-up periods. A particularly promising direction for further investigation is to determine whether the observed

phenomenon of cognitive incubation in M3 can be replicated across other academic disciplines and whether appropriately designed pre-listening activities can effectively shorten this latent period among B1 learners.

In conclusion, the availability of a sophisticated AI tool does not automatically guarantee academic success. The effectiveness of such technology depends on several factors. These include the learners for whom it is designed, the way it is integrated into the instructional process, and the extent to which it engages students in active and deep cognitive processing. As NotebookLM-generated podcasts become an increasingly common component of ESP education, the central question is no longer whether they should be incorporated. Instead, the focus should be on how they can be implemented effectively. Successful AI integration in education requires an approach that is responsive to students' current proficiency levels while systematically providing cognitive challenges that extend slightly beyond their existing capabilities.

Sources

1. ALBRECHT-CRANE, Christa. 2025. Thinking smarter, not harder? Google NotebookLM's misalignment problem in education. In: Proceedings of SIGDOC '25. New York: ACM. [cit. 2025-06-18] ISBN 979-8-4007-0368-6. Available: <https://doi.org/10.1145/3711670.3764628>.
2. ANĐEL, Jiří. 1998. Statistické metody. 2. vyd. Praha: Matfyzpress. ISBN 80-85863-18-2.
3. COHEN, Jacob. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates. ISBN 978-0-8058-0283-2.
4. FIELD, Andy. 2009. Discovering statistics using IBM SPSS statistics. 6th ed. London: SAGE Publications. ISBN 978-1-84787-907-3.
5. GERLICH, Michael. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(6), s. 1–28. [cit. 2025-06-18]. Available: <https://doi.org/10.3390/soc1506xxxx>.
6. HENDL, Jan. 2006. Přehled statistických metod zpracování dat. Praha: Portál. ISBN 80-7367-196-4.
7. KESTIN, Greg; MILLER, Kelly; KLALES, Alexis et al. 2024. AI tutoring outperforms active learning. Research Square [preprint]. [cit. 2025-06-18]. Available: <https://doi.org/10.21203/rs.3.rs-4243877/v1>.
8. KRASHEN, Stephen D. 1985. The input hypothesis: Issues and implications. London: Longman. ISBN 0-582-55384-0.
9. LEE, Hao-Ping; SARKAR, Advait; TANKELEVITCH, Lev et al. 2025. The impact of generative AI on critical thinking. In: Proceedings of the CHI '25 Conference on Human Factors in Computing Systems. New York: ACM. [cit. 2025-06-18]. Available: <https://doi.org/10.1145/3706598.3713778>.
10. LEHMACHER, Wolfgang. 1978. A more general approach to the analysis of clinical trials. Biometrical Journal, 20(3), s. 281–289.
11. PENG, Chao-Ying Joanne; CHEN, Li-Ting. 2015. Beyond Cohen's d: Alternative effect size measures for between-subject designs. Journal of Experimental Education, 82(1), s. 22–50.
12. POLVERINI, Giulio; GREGORCIC, Bor. 2024. How understanding large language models can inform the use of ChatGPT in physics education. European Journal of Physics, 45(2), 025701.
13. REYNA, Jorge. 2025. The potential of Google NotebookLM for teaching and learning. In: eLearn 2025 Proceedings. Phoenix, AZ: AACE.
14. RUBLÍK, František. 2011. Štatistické testy. Bratislava: VEDA. ISBN 978-80-224-1172-1.
15. SWELLER, John. 1988. Cognitive Load During Problem Solving: Effects on Learning. In: Cognitive Science. [cit. 2025-06-18]. Available: https://doi.org/10.1207/s15516709cog1202_4
16. TUFINO, Eugenio. 2025. NotebookLM: An LLM with RAG for active learning and collaborative tutoring. arXiv preprint. [cit. 2025-06-18]. Available: <https://doi.org/10.48550/arXiv.2504.09720>.