Using Generative AI to Evaluate Pre-Service Teachers' Project-Based Learning Designs

Nikola Straková [1], *Jan Válek* [2], *Peter Marinič* [3]

DOI: https://doi.org/10.53349/re-source.2026.is1.a1545

Abstract

This paper presents an innovative application of generative artificial intelligence to the evaluation of project-based learning (PBL) lesson preparation by pre-service teachers of vocational subjects. Using a system prompt grounded in the Gold Standard PBL criteria (Larmer, Mergendoller, & Boss, 2015), the AI analyses project quality in terms of intellectual challenge, authenticity, student voice and choice, reflection, critique and revision, and public product, with particular attention to the development of critical thinking and problem-solving skills. The evaluation process includes criterion-based commentary, identification of strengths and areas for improvement, and the provision of constructive feedback to students. The primary aim is to support the development of key professional competencies among future teachers and to enhance their ability to design meaningful and effective project-based instruction. The paper further discusses the pedagogical benefits, limitations, and ethical considerations associated with this approach.

Keywords: Artificial Intelligence, Project-Based Learning, Lesson Design Evaluation, Gold Standard PBL, Pre-Service Teachers, Vocational Education, Feedback, Didactic Reflection

1 Introduction

Throughout history, humans have sought to improve their living conditions through the development of tools and technologies, primarily aimed at reducing the physical demands of labour. Historical analysis reveals several periods marked by fundamental transformations in production processes and modes of work, commonly referred to as industrial revolutions. The

[1] Masaryk University, Faculty of Education, Dept. of Physics, Chemistry and Vocational Education, Poříčí 7, CZ 60300 Brno, Czech Republic. *E-Mail:* strakova@mail.muni.cz
[2] Masaryk University, Faculty of Education, Dept. of Physics, Chemistry and Vocational Education, Poříčí 7, CZ 60300 Brno, Czech Republic. *E-Mail:* valek@ped.muni.cz
[3] Masaryk University, Faculty of Education, Dept. of Physics, Chemistry and Vocational Education, Poříčí 7, CZ 60300 Brno, Czech Republic. *E-Mail:* marinic@ped.muni.cz

most recent, the Fourth Industrial Revolution, extends beyond automation and robotization to include the integration of advanced information technologies into production systems. In the context of increasing automation and robotization, concerns have emerged regarding potential job displacement (Oosthuizen, 2022). At the time, creative work was generally considered less vulnerable to such changes; however, this assumption has been challenged by the rapid advancement of artificial intelligence. Artificial intelligence is based on people's efforts to describe and simulate learning processes (Zirar, 2023), (Ahmmed et al., 2024). People have long sought to develop various algorithms capable of solving different problems (Zirar et al., 2023). This effort is currently manifested in the form of artificial intelligence, the core of which is based on processes that we may not be familiar with and are therefore unable to describe.

On the one hand, we can view artificial intelligence as something that can threaten us and potentially take away jobs based on creativity. On the other hand, however, if we learn to use artificial intelligence, it can significantly increase the efficiency of our work. Artificial intelligence is here to stay, so we need to learn how to use it and take advantage of it. This also applies to the educational process. (Zirar, 2023), (Zirar et al., 2023), (Ahmmed et al., 2024).

2 Defining and Contextualizing Artificial Intelligence

According to a survey conducted by Ipsos in collaboration with Vodafone, up to 87% of students use AI tools at home or at school. In addition, 74% of students consider it important to have teachers who understand AI and know how to work with it in class. Students use artificial intelligence to make lessons more attractive (33%), to make it easier to work in unpopular subjects (23%), to support learning in favourite subjects (22%), develop skills instead of purely theoretical knowledge (17%), improve grades (16%), individualize teaching (15%), or personalize assignments (10%). (Ipsos, 2025)

The National Pedagogical Institute of the Czech Republic also covers the issue of artificial intelligence within the field of digitization in education. Teachers can find a wealth of information here on how artificial intelligence works, selected legislative aspects of its use in the educational process, how to use artificial intelligence in teaching, and much more, including answers to frequently asked questions (Šnajdrová, 2025). Teachers can also find inspiration on the Artificial Intelligence in Schools Czech web portal (Artificial Intelligence in Schools (AI in Schools)).

Teachers use artificial intelligence as inspiration for their teaching and for creating teaching materials in the form of various teaching resources. However, teachers can also use artificial intelligence to assess students, most often in connection with providing feedback on students' written work. According to the above-mentioned research, 53% of students think that grading using AI is fairer than grading by teachers. On the other hand, 48% of young people fear that if AI is used for assessment, some students may be discriminated against. (Ipsos, 2025)

The use of artificial intelligence for student assessment is therefore considered risky by the students themselves. In addition to them, the EU AI Act (Regulation - EU - 2024/1689 - EN - EUR-Lex) also considers the use of artificial intelligence for assessment to be risky. It thus regulates the use of artificial intelligence for assessment only as a basis for teachers, who will then carry out the final assessment independently. This approach is similar to that mentioned by Harari in his book Nexus regarding the use of artificial intelligence in the judiciary to prepare the basis for judgments. If the use of artificial intelligence is part of such an important process, the people affected by the consequences of this process are entitled to an explanation and review of the results of this process. (Harari, 2025)

In the previous section, we mapped out the broader historical and technological context—from moments of fundamental change in manufacturing to the current phase, which we refer to as the fourth industrial revolution—and then focused on the role of artificial intelligence in the transformation of work, production, and creativity. In this chapter, we will move closer to the school and educational environment: we show how AI technologies are actually reflected in school and student activities, including the use of AI tools, their perception by students and teachers, and the legal and ethical aspects of their use in assessment. (Yue Yim, 2024), (Wang et al., 2024), (Ruiz Viruel et al., 2025)

This transitional phase integrates a macro-level perspective on technological change with a micro-level perspective focused on a specific educational context. The following section presents the research component of the study, outlining the methodology and research design. It builds on the technological and educational insights discussed in the preceding chapters and proposes a framework for examining the application of generative artificial intelligence in the educational process, specifically within project-oriented learning. This approach ensures that the empirical investigation is grounded in both technological and educational theory while remaining relevant to school practice and the professional needs of teachers.

3 Methodology

The primary objective of the methodology is to examine the applicability of a system prompt for generative artificial intelligence (GenAI) in evaluating the preparation of project-based teaching by students of vocational subjects. The study investigates whether GenAI can provide relevant, consistent, and educationally valuable feedback.
Sub-goals:

- Compare how different genAI tools (chatbots) interpret the same prompt and how their outputs differ.
- Compare outputs from genAI with expert (teacher) evaluation.

Research questions:

- Which areas of PBL are best developed in projects?
- What is the degree of agreement between AI evaluation and teacher evaluation?
- How do the outputs of different AI tools differ when using the same prompt?

3.1 The System Prompt for Evaluating Project-Based Teaching Preparations

In the system prompt we have created, genAI expert plays the role of a project teaching expert who specializes in evaluating projects according to the "Gold standard PBL" described in the book by Larmer, Mergendoller & Boss (2015). The expert's goal is to evaluate the project preparations sent by the teacher and provide constructive feedback.
At the beginning of the conversation, the user is prompted to upload the project preparation. GenAI then reads the preparation in detail and evaluates it according to seven criteria corresponding to the seven Gold standard PBL, by Larmer, Mergendoller & Boss (2015):

- Challenging Problem or Question,
- Sustained Inquiry,
- Authenticity,
- Student Voice & Choice,
- Reflection,
- Critique & Revision,
- Public Product.

For each criterion, genAI adds a detailed comment on how well the criterion is implemented in the project preparation (not just whether it is mentioned or not). GenAI will also list the strengths and weaknesses of the project preparation, including recommendations for improvement. Finally, it will classify the project preparation according to one of four levels of PBL gold standard implementation:

1 = Meets the Gold Standard,
2 = Meets with minor reservations,
3 = Partially meets,
4 = Needs significant improvement.

The implementation of the seven Gold Standard PBL ensures that projects will be challenging, engaging, complex, and interesting for students, leading to beneficial educational outcomes. At the same time, according to Larmer, Mergendoller & Boss (2015), the goal of project-based learning is for students to develop not only key knowledge and understanding, but also "21st-century skills" such as critical thinking/problem solving, collaboration, and self-management. The system prompt designed for various genAI conversation tools (chatbots) is designed to consider the theory of project-based learning according to Larmer, Mergendoller & Boss (2015) and to ensure that the resulting preparation and implementation of projects in teaching are of high quality. Evaluating projects before their implementation in real

educational practice allows for the correction of planned projects, thereby contributing to higher quality project-based teaching. GenAI includes a million experts with a million areas of expertise who can supplement the teacher's evaluation with valuable comments.

3.2 System Prompt Testing Process: Input Data, Outputs, Comment Types

The system prompt is tested on a sample of 10 didactic preparations for project-based teaching created by students of practical teaching and specialized subjects as part of a project-based teaching course.

Each PBL plan is evaluated by genAI using the created system prompt. From the available genAI tools, the chatbots ChatGPT, Gemini, Claude, Copilot, and Perplexity were selected. At the same time, the plans were evaluated in parallel by experts—two teachers of the Vocational education, Practical Teaching and Vocational Subjects study program. The evaluation results were compared to determine how different genAI tools interpret the same prompt and how their outputs differ, including a comparison with the output from experts.

The implementation of Gold standard PBL (Larmer, Mergendoller & Boss, 2015) was evaluated on a scale of 1 to 4:

1 = Meets the Gold Standard – Criteria are fully and exceptionally met; the project is inspiring and demonstrates deep understanding;
2 = Meets with minor reservations – Criteria are mostly met, but there are small shortcomings or opportunities for improvement;
3 = Partially meets – Criteria are only partially met, significant improvement is required;
4 = Needs significant improvement – Criteria are not met or are minimally met; the project requires major revision.

The strengths and weaknesses of the projects were evaluated verbally.

3.3 Methodology of Verbal Assessment Analysis

The verbal assessment analysis procedure is shown in Figure 1 and began with the collection of all verbal assessments into an MS Excel spreadsheet for simple tabular coding, where the source of the assessment (genAI vs. teacher) was noted.

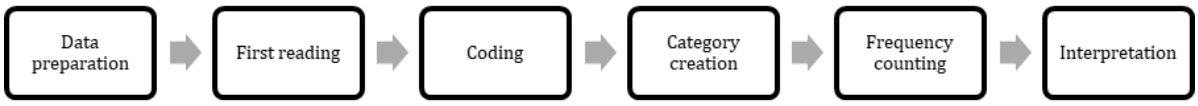

Figure 1: Word evaluation analysis procedure.

After thoroughly reading the evaluations, sentences or parts of sentences with the same meaning were identified and a code was assigned to these units of meaning. The codes were

created inductively (they gradually emerged from the data/text of the evaluation). The codes were then sorted by similarity into 12 semantic categories, see Table 1.
The assessments were then reread, this time with the aim of determining how many assessments fell into each category (frequency) and whether the assessment was listed as a strength or weakness. Frequencies were recorded separately for individual evaluators (genAI vs. teacher) and by group (strengths vs. weaknesses).
The data was organized into a clear frequency table and supplemented with excerpts from the evaluations as illustrations for each category. The resulting data was then interpreted in terms of existing contrasts between evaluators and the most common strengths and weaknesses.

Codes	Category
Real world, practice, authenticity, community, real problem, client	1. Authenticity and connection to the real world
Professional context, professional skills, professional profiling, practice, professional standards	2. Professional relevance and professional development
Interconnection of fields, interdisciplinarity, integration of theory and practice	3. Interdisciplinarity and interconnection of areas
Autonomy, freedom of decision-making, creativity, own ideas	4. Student autonomy and creativity
Social impact, community, civic dimension, social responsibility	5. Social and community impact
Clear structure, good management, quality planning, criticism & revision, Gold Standard	6. Quality of pedagogical design
Reflection, feedback, evaluation, critical thinking	7. Reflection, critical thinking, and evaluation
Student involvement, motivation, enthusiasm, engagement	8. Student motivation and engagement
Lack of research, superficial work, instructions, little discovery	9. Weak inquiry process
No output, non-public project, weak presentation	10. Weak public product (public output)
(Non)demanding in terms of materials, teaching aids	11. Demanding material requirements
Great, excellent, no comments	12. Overall evaluation: excellent

Table 1: Conversion of code to semantic categories.

4 Results

The point scores according to the seven Gold Standard PBL criteria (C1–C7), including the strengths and weaknesses of the projects, are shown in Table 2 (teachers did not explicitly evaluate the strengths and weaknesses for each criterion, but only for the project as a whole). Of all the AI tools, the chatbot Gemini (mean value 1.36 points) received the most positive evaluation (1 = Meets the Gold Standard – Criteria are fully and exceptionally met, the project

is inspiring and demonstrates deep understanding), followed by Perplexity (mean value 1.46 points) and Copilot (mean value 1.75).
On the other hand, the most critical were the Claude chatbot (mean value is 2.50 points) and ChatGPT (mean value 1.88 points). Claude and ChatGPT evaluated all criteria with reservation and saw room for improvement in each (especially in criteria C5 Reflection and C6 Criticism and Revision).
The average overall rating from all AI chatbots was the same as the average overall rating from the two teachers (mean value 1.79 points).

	ChatGPT	Copilot	Gemini	Perplexity	Claude	genAI average	Teacher
C1 - Challenging Problem or Question	1,9	1,4	1,3	1,8	3,0	1,88	1,6
C2 - Sustained Inquiry	2,3	2,3	1,1	1,5	2,9	2,02	2,05
C3 - Authenticity	1,0	1,0	1,0	1,0	2,1	1,22	1,35
C4 - Student Voice & Choice	2,0	2,0	1,5	1,5	2,6	1,92	1,65
C5 - Reflection	2,7	2,4	1,6	2,1	3,3	2,42	2,35
C6 - Critique & Revision	2,8	2,6	2,2	2,1	3,1	2,56	2,05
C7 - Public Product	1,3	1,5	1,7	1,1	2,0	1,52	1,45
Strengths and weaknesses for each criterion? (1 - yes/0 - no)	1,0	0,8	0,44	0,6	1,0	0,77	X
Average rating (1 - 4)	**1,88**	**1,75**	**1,36**	**1,46**	**2,5**	**1,79**	**1,79**

Table 2: Results of the analysis of the average point rating of PBL gold standards.

All evaluators, including teachers, agree that the best-fulfilled criterion in student projects is C3 Authenticity. This means that the projects are realistic and connected to professional practice and the local community.
Overall, the criteria C7 Public Product and C1 Challenging Problem or Question are also positively evaluated. This means that the projects have a clearly defined public output (often a presentation or exhibition) and are defined at the outset by a meaningful question or address a real problem (with the exception of the chatbot Claude, which was the only one to rate this criterion as insufficient = 3 points, criticizing the absence of an explicit formulation of the central question of the project).
The weaknesses of the projects across all evaluators are criteria C6 Critique & Revision, C5 Reflection, and C2 Sustained Inquiry. According to the evaluation, the projects are short, without in-depth research, mostly lacking a systematic feedback cycle, and the reflection is formal, without real self-reflection by the students.

The differences between the AI and teacher evaluations can be interpreted as teachers tending to evaluate slightly more positively. The average AI evaluation score ranges from 1.36 to 2.50 points, while teachers evaluate an average of 1.79 points. Overall, however, a comparison of the project scores shows that AI can serve as a valid first stage of evaluation. AI provides comparable assessment results to those of teachers. By combining both approaches, a comprehensive evaluation of PBL preparations can be obtained. AI can quickly and consistently assess the formal aspects, while teachers provide a deeper pedagogical interpretation.

4.1 Results of the Verbal Assessment Analysis

Table 3 summarizes how genAI and teachers describe the strengths and weaknesses of student projects in individual semantic categories. For each category, it shows the proportion of positive and negative mentions and their ratio. Table 3 shows where the genAI and teacher evaluations agree and where they differ.

Category/Evaluator	GenAI			Teachers		
	Number of strong ratings (+)	Number of weak ratings (-)	Ratio (+/–)	Number of strong ratings (+)	Number of weak ratings (-)	Ratio (+/–)
Authenticity and connection to the real world	**96%**	0%	96%	**55%**	20%	35%
Professional relevance and professional development	**70%**	0%	70%	**20%**	0%	20%
Interdisciplinarity and connection between areas	60%	0%	60%	10%	0%	10%
Student autonomy and creativity	34%	48%	-14%	15%	20%	-5%
Social and community impact	26%	0%	26%	10%	0%	10%
Quality of pedagogical design	46%	8%	38%	0%	0%	0%
Reflection, critical thinking, and evaluation	20%	**82%**	-62%	20%	**60%**	-40%
Student motivation and engagement	20%	**48%**	-28%	0%	0%	0%
Weak inquiry process	0%	**40%**	-40%	0%	**45%**	-45%
Weak public product (public output)	0%	30%	-30%	0%	5%	-5%
Material resource requirements	0%	0%	0%	5%	5%	0%
Overall rating: excellent	0%	0%	0%	30%	0%	30%

Table 3: Comparison of strengths and weaknesses identified in project evaluations.

Higher percentages in the genAI columns indicate that chatbots tend to identify and explicitly name more strengths (and weaknesses) than teachers and overall evaluate projects more optimistically and "generously" (in the case of strengths) and more strictly (in the case of weaknesses). Teachers, on the other hand, are more selective, i.e., they highlight only those

aspects that they consider truly significant, which is why their verbal assessments are reflected in lower percentages in the table.

The verbal evaluation confirms the conclusions of the point evaluation. The projects are particularly strong in relation to reality and professional practice (Gen AI marked as a strength in 96% and 70% of evaluations, respectively, and by teachers in 55% and 20%, respectively). On the contrary, the most problematic areas from the perspective of both evaluators are reflection and the process of inquiry (GenAI marked as a weakness in 82% and 40% of evaluations, respectively, and by teachers in 60% and 45% of evaluations, respectively). GenAI also mentions a frequent lack of motivation and internal engagement (48%). The teacher does not comment on this area.

An interesting difference emerges in what the individual evaluators focus on. GenAI places relatively strong emphasis on "Quality of pedagogical design" (positive ratio +38%) and systematically identifies the formal qualities of projects, while the teacher does not single out this category and focuses more on the process and impact of learning. In addition, the teacher is the only one to include the category "Overall rating: excellent" (+30%), which shows that, in addition to the sub-criteria, he also considers the holistic impression of the project, its educational potential, including an assessment of the students' overall work.

4.2 Nature of the Evaluations Generated by Individual Chatbots

All tested chatbots evaluated projects according to a similar pattern: for each criterion, they described the status of the project preparation, listed its strengths and weaknesses, and in most cases added a final summary, including the level of fulfilment of the Gold Standard PBL.

ChatGPT produced concise, factual, and quickly generated evaluations (approximately 2–3 pages) that combined a description, pros, cons, and recommendations for each criterion and concluded with a clear summary of strengths and weaknesses.

Claude produced the longest and most detailed texts (4–8 pages), with extensive elaboration on strengths and weaknesses, but at the cost of significantly longer generation times.

Copilot provided more concise, clear evaluations (1.5–2 pages), with a clear summary of the criteria and a brief final overview; recommendations were rather rare.

Gemini combined factual comments with visual clarity (bold highlighting, bullet point summaries), and the outputs were of medium length (approx. 2.5–3 pages) and quickly generated.

Perplexity opted for compact paragraphs combining description, strengths and weaknesses, and recommendations, concluding with a brief summary and Gold Standard PBL level (2–3 pages, quick generation).

5 Conclusion

A comparison of generative artificial intelligence (GenAI) and teachers indicates that AI is well suited for rapid, structured assessment, whereas teachers remain essential for evaluating the depth of pedagogical processes and student development. GenAI demonstrates sensitivity to formal aspects of project quality, such as design, authenticity, and relevance, while teachers are better positioned to capture procedural and pedagogical dimensions, including the learning process and the extent of student engagement in project preparation. The integration of both perspectives therefore provides a multidimensional understanding of the quality of project-based teaching preparation. GenAI can function as a tool for efficient, criteria-based evaluation aligned with the Gold Standard PBL framework, while teachers contribute indispensable interpretative insight into how projects support student learning in practice.

All examined chatbots were able to provide structured evaluations consistent with the Gold Standard PBL criteria; however, they differed in the level of detail, length, and emphasis of their recommendations. For practical application in school settings, tools that balance clarity, conciseness, and actionable feedback—particularly Gemini, ChatGPT, Copilot, and Perplexity—appear most suitable. In contrast, systems producing excessively lengthy and time-consuming outputs, such as Claude, may be less practical for routine use.

References

Ahmmed, M. S., Isanaka, S. P., & Liou, F. (2024). Promoting Synergies to Improve Manufacturing Efficiency in Industrial Material Processing: A Systematic Review of Industry 4.0 and AI. *Machines*, *12*(10), 681. https://doi.org/10.3390/machines12100681

Harari, Y. N. (2025). *Nexus: A Brief History of Information Networks from the Stone Age to AI*. Vintage Publishing.

IPSOS. (2025). *Tisková zpráva: 9 z 10 dětí používá AI, známkovaní od AI však vnímají rozporuplně*: [9 out of 10 children use AI, but they have mixed feelings about AI grading]. IPSOS. https://www.ipsos.com/sites/default/files/ct/news/documents/2025-01 .

Larmer, J., Mergendoller, J., & Boss, S. (2015). Setting the standard for project-based learning: *A proven approach to rigorous classroom instruction.* ASCD.

Oosthuizen, R. M. (2022). The Fourth Industrial Revolution – Smart Technology, Artificial Intelligence, Robotics and Algorithms: Industrial Psychologists in Future Workplaces. *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.913168

Ruiz Viruel, S., Sánchez Rivas, E., & Ruiz Palmero, J. (2025). The Role of Artificial Intelligence in Project-Based Learning: Teacher Perceptions and Pedagogical Implications. *Education Sciences*, *15*(2), 150. https://doi.org/10.3390/educsci15020150

Šnajdrová, L. (2025, May 23). Neděláme z dětí ajťáky [We don't turn children into IT specialists]. https://digitalizace.rvp.cz/clanky/nedelame-z-deti-ajtaky. Národní pedagogický institut České republiky. Retrieved November 16, 2025, from https://digitalizace.rvp.cz/clanky/nedelame-z-deti-ajtaky

Wang, X., Chen, M., & Chen, N. (2024). How artificial intelligence affects the labour force employment structure from the perspective of industrial structure optimisation. *Heliyon*, *10*(5), e26686. https://doi.org/10.1016/j.heliyon.2024.e26686

Yue Yim, I. H. (2024). A critical review of teaching and learning artificial intelligence (AI) literacy: Developing an intelligence-based AI literacy framework for primary school education. *Computers and Education: Artificial Intelligence*, 7, 100319. https://doi.org/10.1016/j.caeai.2024.100319

Zirar, A. (2023). Can artificial intelligence's limitations drive innovative work behaviour? *Review of Managerial Science*, *17*(6), 2005-2034. https://doi.org/10.1007/s11846-023-00621-4

Zirar, A., Ali, S. I., & Islam, N. (2023). Worker and workplace Artificial Intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, 124, 102747. https://doi.org/10.1016/j.technovation.2023.102747

