INSTRUCTIONAL DESIGN IN TEACHER EDUCATION AND GENERATIVE AI: A TWO-YEAR EXPLORATORY STUDY ON THE ALIGNMENT OF AI-GENERATED LEARNING ACTIVITIES WITH THE ANDERSON AND KRATHWOHL TAXONOMY

Ludovica Rizzo
University of Salento & University of Macerata, Italy
ludovica.rizzo3@unisalento.it

Stefania Pinnelli
University of Salento, Italy
stefania.pinnelli@unisalento.it

Double Blind Peer Review

Citation

Rizzo, L., & Pinnelli, S. (2026). Instructional design in Teacher Education and Generative AI: a two-year exploratory study on the alignment of AI-Generated learning activities with the Anderson and Krathwohl Taxonomy. *Italian Journal of Health Education, Sport and Inclusive Didactics*, *10*(3). https://doi.org/10.32043/gsd.v10i3.1825

Doi: https://doi.org/10.32043/gsd.v10i3.1825

Copyright notice:

© 2026 The Author(s). This is an open access, peer-reviewed article published by Cares S.r.l. – REN Press and distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

renpress.it/index.php/gsdj
ISSN: 3103-5744

ABSTRACT

The Anderson and Krathwohl Taxonomy (2001) provides a key framework for differentiated instructional design, supporting the development of learning tasks that move beyond the implicit model of the average student. However, limited empirical attention has been devoted to the alignment between generative AI outputs and the cognitive demands defined by the taxonomy. This exploratory study, conducted over the 2024–2026 biennium, examines how 134 pre-service teachers, organised into 32 working groups, evaluated the taxonomic coherence of learning activities generated by ChatGPT, Gemini and Meta AI across the six cognitive process levels. The findings reveal differentiated patterns of alignment across systems, cohorts and cognitive levels. Rather than positioning generative AI as a substitute for teachers' instructional design competence, the results highlight its potential role as a scaffolding reflective mediator in initial teacher education.

KEYWORDS

Generative AI; Instructional Design; Revised Bloom's Taxonomy; initial teacher education; taxonomic alignment.

Received 2026-06-30
Accepted 2026-10-05
Published 2026-10-30

1. Introduction

The increasing complexity of contemporary classrooms has made it progressively less adequate to rely on instructional models implicitly designed around the profile of a hypothetical average student. Educational contexts are instead characterised by significant intra- and inter-individual variability, which requires teaching practices capable of responding to the heterogeneity of students' cognitive, social, and emotional profiles (d'Alonzo, 2016; Rose, 2016; Tomlinson, 2014). From this standpoint, individual differences should not be regarded as exceptions to be compensated for, but as constitutive dimensions of educational design. Differentiation thus emerges as a fundamental principle for broadening learning opportunities and fostering the development of each learner's potential. Within this framework, the Taxonomy of Learning Objectives, in the revised version proposed by Anderson and Krathwohl (2001), offers a particularly relevant conceptual and operational framework for differentiated instructional design. By organising learning objectives along a continuum of increasingly complex cognitive processes, the taxonomy enables teachers to design learning pathways that are responsive to different levels of prior knowledge, cognitive development, and learning readiness. Its educational value lies not only in the classification of objectives, but also in its capacity to guide the intentional construction of tasks that support students in working within their zone of proximal development. The intent of the taxonomy therefore extends beyond the mere achievement of minimum learning objectives, promoting an approach intentionally oriented towards the development and use of higher-order thinking processes. In this scenario, the rapid development of generative Artificial Intelligence systems has contributed to positioning AI as a potential resource for instructional design (Choi et al., 2024; Karaman & Göksu, 2024; Weng et al., 2024). These systems may support teachers by intervening on some of the variables that often hinder the systematic implementation of differentiated teaching practices, including limited time, availability of resources, and organisational workload, thereby helping to reduce cognitive and organisational overload (Deunk et al., 2018). Moreover, it is increasingly presented as a tool capable of assisting teachers in the development of learning activities aligned with the different levels of Anderson and Krathwohl's taxonomy (Eideh, 2026; Almatrafi & Johri, 2025), potentially supporting a closer correspondence between students' educational needs and the cognitive demands embedded in instructional tasks (Akdeniz et al., 2025). In light of these premises, it seems essential to investigate the extent to which AI-generated activities effectively reflect the cognitive processes associated with the different taxonomic levels, as well as to clarify the role of the teacher in the processes of evaluation, interpretation and critical reworking

of these outputs, in order to ensure their pedagogical and educational consistency. Accordingly, this study examines the degree of alignment between learning activities generated by generative AI systems and the cognitive process levels outlined in Anderson and Krathwohl's taxonomy. By focusing on the evaluations provided by pre-service teachers, the study also aims to highlight the centrality of professional competence in mediating the use of AI in instructional design. In this perspective, AI is not considered a substitute for teachers' pedagogical judgement, but rather a potential reflective tool through which future teachers can learn to analyse the cognitive demands of tasks and ensure that learning activities authentically support the educational success of all students.

2. Theoretical framework

2.1 Teachers' instructional competence in the design of personalised tasks and the integration of AI

Contemporary higher education is progressively oriented towards a reconfiguration of its pedagogical and assessment paradigms, moving beyond teacher-centred transmission models in favour of student-oriented approaches that emphasise the development of transferable skills, the activation of higher-order thinking processes and forms of active participation. Within this transformation, assessment practices are increasingly called upon to transcend the mere measurement of acquired knowledge and consider students' ability to mobilise knowledge and skills in response to complex and authentic situations, integrating key 21st-century skills such as critical thinking, problem solving, creativity, and collaboration (Mendis et al., 2025). This orientation is explicitly reflected in the main international reference frameworks dedicated to the development of teachers' professional skills in relation to the use of artificial intelligence. In particular, the UNESCO AI Competency Framework for Teachers (UNESCO, 2024) emphasises the need for teachers of all levels to develop teaching skills geared towards the conscious integration of AI in the design and facilitation of student-centred learning practices. The stated objective is to promote active student engagement, support differentiated teaching and improve the quality of educational interactions. Central to this framework, and explicitly referred to therein, is the ability to design learning tasks that are sensitive to student profiles, capable of activating different types of knowledge and stimulating higher-order cognitive processes, in line with the conceptual framework of the Taxonomy proposed by Anderson and Krathwohl.

In this context, the role of the teacher is significantly broader than just critically analysing the output generated by AI systems or monitoring phenomena such as bias, hallucinations and distortions, which can lead to outcomes that diverge from the stated educational intentions (Kasneci et al., 2023). It includes the collection, selection and conscious use of data as a basis for designing personalised learning paths that are functional to the individual needs and profiles of students (Farhood et al., 2025). In addition, the practical use of AI in the design of personalised learning tasks is still predominantly addressed in research from a technological rather than a pedagogical-didactic perspective, often favouring technical efficiency at the expense of educational intent and the actual effectiveness of the teaching-learning processes activated (Zawacki-Richter et al., 2019). This highlights the need to anchor AI-supported learning practices more firmly in pedagogical frameworks that can guide instructional design towards authentic personalisation of learning, functional to the development of individual potential. In this context, generative AI is emancipated from its role as a mere technical tool supporting teaching practice to become a pedagogical artefact, whose educational value is realised through the reflective processes of analysis, evaluation and critical re-elaboration by teachers. Involving teachers in initial training in the assessment of the alignment between AI-generated learning tasks and the cognitive process levels outlined by Anderson and Krathwohl's Taxonomy means therefore, enhancing their ability to critically examine the cognitive demands implicit in the proposed tasks, to identify ambiguities between contiguous taxonomic levels and to unmask higher-order tasks that in reality only simulate complexity without substantially activating deeper cognitive processes. Recent literature points to the need to move beyond a conception of AI as a static tutoring tool and to frame it instead as an adaptive system capable of modulating the granularity of its support – that is, the degree of detail and precision of its outputs – according to the user's level of uncertainty, thereby enabling a process of cognitive co-regulation (Luckin, 2018). From this perspective, AI-mediated support may be understood as a form of algorithmic scaffolding that operates within the individual's zone of proximal development, providing calibrated prompts and suggestions that can act as catalysts for cognitive restructuring (Van der Graaf et al., 2023).

2.2 Bloom's Taxonomy revised by Anderson and Krathwohl

The taxonomy of learning objectives in the revised version proposed by Anderson and Krathwohl (2001) provides teachers with a dynamic structure

that allows them to relate learning objectives – and the teaching tasks associated with them – to the cognitive processes and skills necessary for their achievement. By articulating learning along two interconnected dimensions, knowledge and cognitive process, the taxonomy allows teachers to adapt their teaching to the potential of their students, progressively supporting deeper forms of understanding and increasingly sophisticated ways of thinking. Within this model, knowledge is conceptualised as the "what" of learning and is organised along a continuum that develops from more concrete to progressively more abstract forms. The knowledge dimension comprises four distinct categories. Factual knowledge concerns the basic elements of a discipline, such as terminology and specific details, and represents the fundamental "products" of learning. Conceptual knowledge includes the relationships between these elements and allows them to function coherently within broader structures; it includes classifications, categories, principles, generalisations, theories, models and conceptual frameworks, also referring to the "what" of learning, but at a higher level of abstraction. Procedural knowledge concerns knowing how and why certain activities are performed, including methods of investigation, skills, algorithms, techniques, and discipline-specific procedures; it therefore corresponds to the procedural dimension of the "how" of learning. Finally, metacognitive knowledge involves awareness of one's own cognitive processes and includes knowledge of cognitive strategies, learning processes, and the conditions that promote their effectiveness. The cognitive processes dimension, on the other hand, comprises six levels organised along a continuum of increasing complexity, developing from lower-order to higher-order cognitive processes. Each level is further divided into specific cognitive processes, which are useful for clearly defining learning expectations for students. In particular, the Remember level concerns the retrieval of relevant information from long-term memory; Understand involves determining the meaning of educational materials, including oral, written and graphic representations; Apply concerns the execution or use of a procedure in a given situation; Analyse consists of breaking down the material into its constituent parts and determining their relationships with each other and with the overall structure; Evaluate involves expressing judgements based on criteria and standards – quantitative or qualitative – relating, for example, to quality, effectiveness, efficiency or consistency; finally, Create involves reorganising or assembling elements in order to produce a model or structure that was not immediately present before.

This two-dimensional structure allows teachers to implement a teaching plan that goes beyond tasks based exclusively on retrieving information from long-term memory – still widespread in many school contexts – and moves towards activities that promote meaningful interpretation, in-depth processing and active knowledge construction. As Anderson (2005) points out, effective learning depends on the degree to which new information is processed, integrated and transformed, rather than simply passively accumulated.

1.3 Aims of the study

Based on the theoretical framework outlined above, the present study pursues the following aims: (1) to examine the degree of alignment between learning activities generated by artificial intelligence systems and the cognitive process levels delineated in the Anderson and Krathwohl taxonomy; (2) to identify which cognitive process levels present greater criticalities or ambiguities in the evaluation of AI-generated learning activities; (3) to explore qualitative differences in the evaluation of learning activities generated by different generative AI platforms.

These objectives are addressed through the following research questions:

RQ1: To what extent are learning activities generated by artificial intelligence aligned with the cognitive process levels specified in the Anderson and Krathwohl taxonomy?

RQ2: Which cognitive process levels emerge as the most problematic and/or ambiguous in the evaluation of AI-generated learning activities?

RQ3: Are there identifiable qualitative differences in the evaluation of learning activities generated by different artificial intelligence systems?

3. Methodology

3.1 Participants

The participants in the study were students enrolled in the fourth year of the degree programme in Primary Teacher Education, organised into small working groups. All participants had previously completed foundational courses in pedagogy and didactics, as well as specific training on the cognitive processes defined in Anderson and Krathwohl's taxonomy. In the first cohort, 92 students initially took part in the activity. After the exclusion of records that did not meet the inclusion criteria or were not fully compliant with the research procedure –

for example, cases in which groups selected more than one AI system instead of identifying a single preferred output – 46 students were retained for the final analysis, organised into 13 working groups. In the second cohort, 109 students initially participated. Following the application of the same inclusion criteria, 88 students were included in the final dataset, organised into 19 working groups. Overall, the final sample consisted of 134 pre-service teachers and 32 working groups. Each group was asked to design, analyse and evaluate learning activities intentionally aligned with the six levels of the cognitive process dimension of Anderson and Krathwohl's taxonomy. The activities were designed with reference to teaching contexts in the fourth and fifth years of primary school. The collaborative working format was adopted to encourage discussion, argumentative justification and reflective analysis of the instructional decisions involved in the task.

3.2 Study design

The study is situated within an Instructional Design Case Studies approach, focusing on the systematic analysis of instructional design artefacts developed by pre-service teachers with the support of generative artificial intelligence. The procedure was organised into sequential phases.

Each group worked from a previously defined classroom context and was asked to identify a single target ability and the corresponding competence, formulated as an observable learning outcome in terms of what the student would be able to do. On this basis, each group developed a structured prompt aimed at generating learning activities corresponding to the six levels of the cognitive process dimension of the Anderson and Krathwohl taxonomy: *Remember, Understand, Apply, Analyse, Evaluate* and *Create*. The same prompt was submitted to three generative AI systems: ChatGPT, Gemini and Meta AI. The use of the same prompt across systems was intended to reduce variability attributable to the formulation of the request and to make the outputs more directly comparable.

Second, the outputs generated by the three systems were collected and organised according to the corresponding taxonomic level.

Third, the groups of students selected the AI system that, in their perspective, had generated the most appropriate activity for each taxonomic level. In the second cohort, students also rated the perceived taxonomic coherence of each output using a five-point Likert scale, whereas in the first cohort they provided qualitative comments on the outputs generated by the three AI systems.

Finally, the researchers reviewed the outputs and validated the final attribution. The final dataset included 192 taxonomic-level units of analysis, derived from the activities produced by the 32 working groups across the six cognitive process levels. For each unit, three AI-generated outputs were considered, one for each system. Overall, 576 AI-generated activities were analysed, corresponding to the outputs produced by the three systems for each taxonomic-level unit. These activities were subsequently subjected to binary alignment coding by the researchers. Rather than measuring agreement between students and researchers, the final attribution reflected a validation process in which students' selections were reviewed against the researchers' alignment coding. Students' choices were used as a starting point for identifying the output perceived as most appropriate for each taxonomic level. When the selected activity was also judged by the researchers to be sufficiently aligned with the intended cognitive process, the attribution was confirmed. When the selected output did not meet this criterion, the researchers reconsidered the three alternatives and assigned the final attribution to the activity that best approximated the expected taxonomic demand. Where no output provided a sufficiently coherent operationalisation of the intended cognitive process, the case was coded as 0, indicating that none of the systems had generated an adequately aligned activity.

3.3 Data analysis

The analysis combined descriptive quantitative procedures and theory-driven qualitative interpretation. First, the binary alignment scores assigned by the researchers were analysed by cohort, AI system and taxonomic level. Second, the final validated outcome was examined in order to identify which system most frequently produced the output considered closest to the intended cognitive process, as well as the proportion of cases in which no system produced an adequately aligned activity. Third, for the second cohort, students' Likert ratings were analysed to explore perceived taxonomic coherence across the three systems (1 = not coherent at all; 2 = slightly coherent; 3 = partially coherent; 4 = coherent; 5 = fully coherent). The scale was designed to capture the extent to which the activity required, in a central and non-avoidable way, the cognitive process corresponding to the declared taxonomic level.

The qualitative interpretation focused on recurrent discrepancies between the declared taxonomic level and the cognitive demand actually embedded in the task, with particular attention to activities formally associated with higher-order thinking but lacking the expected depth of cognitive engagement.

4. Results

4.1 Alignment of AI-Generated learning activities

The analysis reveals a non-uniform distribution of alignment between the learning activities generated by artificial intelligence systems and the levels of the cognitive processes articulated in the Anderson and Krathwohl taxonomy. Across all three systems analysed, instructional proposals that elicit basic or moderately complex cognitive operations tend to be positioned more precisely within the declared taxonomic level, whereas activities oriented toward higher-order cognitive processes more frequently display incongruence between the formal demand and the cognitive operations actually activated. The convergence of this evidence across evaluation groups, regardless of the system considered, strengthens the assumption that the observed misalignment cannot be attributed to contingent factors, instead suggesting the presence of a structural criticality.

With regard to basic cognitive levels, activities generated for *Remember* and *Understand* were, in most cases, evaluated as fully consistent with their respective taxonomic categories. Evaluations converged in indicating that such tasks predominantly activate processes of retrieval, recognition, and explicit understanding of information, without requiring operations of conceptual re-elaboration or complex integration. Qualitative comments highlight that these activities are characterized by a linear structure and by an immediately recognizable cognitive demand, which facilitates their attribution to the appropriate taxonomic level. In this domain, artificial intelligence systems demonstrate substantial effectiveness in generating tasks of low cognitive complexity, for which distinctions among taxonomic levels are relatively explicit and less prone to interpretative ambiguity.

Greater variability emerges in relation to the intermediate cognitive processes of *Apply* and *Analyse*. Although many activities attributable to these levels were nonetheless judged as aligned, in several cases tasks associated with *Analyse* primarily required the execution of familiar procedures or superficial comparisons, without an explicit decomposition of elements or an examination of relationships among parts. A recurring pattern concerned the tendency to conflate *Analyse* with *Understand*, and, more specifically, to confuse *comparison* with *distinction*. While *comparison* requires the identification of similarities and differences between two or more objects, events, ideas, problems, or situations, the analytic sub-process of *distinguishing* implies the

ability to identify the constituent parts of a broader structure and to determine their relevance, function, or relative importance within that structure. In this sense, several AI-generated activities appeared to remain at the level of descriptive comparison or basic understanding, rather than activating the more specific cognitive operations required by the *Analyse* level. These ambiguities frequently led to judgments of partial alignment, highlighting a blurred boundary between contiguous taxonomic levels and a not always adequate operationalization of the cognitive demand in AI-generated activities.

The *Evaluate* level showed a more variable pattern. In the first cohort, activities presented as evaluative often lacked explicit criteria, standards, or parameters for judgment, thereby reducing the evaluative task to the mere expression of personal opinions. In the second cohort, however, this level showed higher alignment values, suggesting a stronger tendency among AI systems to include explicit criteria, standards, or judgment parameters within activities designed to target the Evaluate level.

The most critical results concerned the higher-order cognitive process of *Create*. Across all AI systems considered, this level was associated with the highest frequency of misalignment and elicited the most articulated qualitative justifications. Tasks classified as creative were frequently formulated as open-ended activities requiring simple retrieval and juxtaposition of information, without clear constraints or design requirements. As a consequence, such activities were perceived as inauthentic higher-order tasks: formally sophisticated and nominally attributed to the *Create* level, yet unable to activate, in a rigorous and conceptually grounded manner, the cognitive processes envisaged by the taxonomy. Particularly emblematic in this respect are tasks framed as “create a poster on the topic,” which, while formally classifiable as creative activities, tend to privilege the aesthetic–expressive dimension of production at the expense of explicit demands for cognitive synthesis, conceptual integration, and restructuring of content.

4.2 Pre-service teachers’ quantitative evaluations of AI-generated learning activities

According to the evaluations provided by the pre-service teachers, Gemini was the system that most frequently generated the activity perceived as most appropriate for the intended taxonomic level. Across the two cohorts, Gemini was selected in 88 cases out of 192, corresponding to 45.8% of the total.

ChatGPT was selected in 53 cases, corresponding to 27.6%, while Meta AI was selected in 51 cases, corresponding to 26.6%.

In the second cohort, students also rated the perceived taxonomic coherence of the outputs on a five-point Likert scale. The mean scores show that Gemini was perceived as the most coherent system, with an average score of 4.31. Meta AI obtained an average score of 3.92, while ChatGPT obtained an average score of 3.66.

Rather than measuring agreement between pre-service teachers and researchers, the final attribution reflected a validation process in which students' selections were reviewed in light of the researchers' alignment coding. In 154 cases out of 192, the initial selections made by the pre-service teachers were retained because the selected output had been coded by the researchers as sufficiently aligned with the intended taxonomic level. In the remaining 38 cases, the selection was revised because the chosen system had not produced an adequately aligned output. In these instances, the researchers re-examined the three outputs and identified the activity that showed the closest correspondence with the expected cognitive demand.

Following this validation process, Gemini remained the system most frequently associated with a final positive attribution, with 74 cases out of 192, corresponding to 38.5%. ChatGPT was identified in 58 cases, corresponding to 30.2%, while Meta AI was identified in 43 cases, corresponding to 22.4%. In 17 cases, corresponding to 8.9% of the dataset, none of the three AI systems produced an output judged sufficiently aligned with the intended taxonomic level.

Table 1: Final validated attribution of students' preferred outputs.

Final validated outcome	*First cohort*	*Second cohort*	*Total*
ChatGPT	43.6%	21.1%	30.2%
Gemini	34.6%	41.2%	38.5%
Meta AI	14.1%	28.1%	22.4%
No adequately aligned output	7.7%	9.6%	8.9%

In addition, as highlighted by the validation process, the findings point to the still-developing nature of pre-service teachers' professional skills in analysing,

decomposing, and operationalising cognitive processes within instructional tasks. In many cases, in fact, activities that initially appeared to be aligned – particularly at the *Analyse*, *Evaluate* and *Create* levels – revealed, upon closer analysis, conceptual ambiguities that were potentially relevant to students, especially in relation to the clarity of the cognitive request and the readability of the expected mental operations. Attention to the cognitive processes activated, as distinct from a focus solely on the final product, requires a level of design awareness that is not immediately accessible in the early stages of engagement with cognitively complex tasks, partly because of established school practices that tend to favour outcomes over processes, and lower-order thinking processes over higher-order ones.

From this perspective, AI-generated learning activities take on an educational value that transcends their instrumental function, becoming reflective tools for the critical analysis of educational design. Their use has made it possible to highlight both the strengths and weaknesses in the thought processes that guide future teachers in setting assignments, promoting greater awareness of the distinction between the formal complexity of the task and the cognitive complexity actually required.

4.3 Qualitative differences in pre-service teachers' evaluations of activities generated by different AI systems

Although the overall pattern of results appears substantially homogeneous across the different artificial intelligence systems analysed, the evaluations provided by the first cohort highlighted relevant qualitative differences in the configuration of the generated tasks. In particular, ChatGPT was described as being more strongly oriented toward the structural coherence of activities and toward an explicit internal organization of the task, features that facilitated the identification of the predominantly elicited cognitive process.

By contrast, Gemini was frequently associated with greater expressive richness and a higher potential for student engagement, but also with a tendency to combine heterogeneous cognitive demands within the same task. This integration, while at times perceived as didactically stimulating, increased the level of interpretative ambiguity with respect to taxonomic alignment.

The outputs produced by Meta AI were more often evaluated as generic or redundant, especially with reference to higher-level cognitive processes. In these cases, the groups noted a lower degree of articulation of the cognitive demand and a limited explicitness of objectives, criteria, and task constraints –

elements that negatively affected the taxonomic robustness of the proposed activities.

4.4 Taxonomic alignment of AI-generated activities across systems, cohorts and cognitive levels

The alignment coding conducted by the researchers revealed a non-uniform distribution of taxonomic coherence across AI systems and cohorts. Overall, Gemini achieved the highest proportion of aligned outputs, with 154 aligned activities out of 192, corresponding to 80.2% of the total. Meta AI produced 132 aligned outputs, corresponding to 68.8%, while ChatGPT produced 129 aligned outputs, corresponding to 67.2%.

However, the comparison between the two cohorts indicates that this pattern was not entirely stable across the two years considered. In the first cohort, ChatGPT and Gemini obtained the same proportion of aligned outputs, both reaching 74.4%, whereas Meta AI reached 65.4%. In the second cohort, by contrast, Gemini showed a more pronounced advantage, with 84.2% of aligned outputs, compared with 71.1% for Meta AI and 62.3% for ChatGPT. This result suggests that Gemini's overall advantage in terms of binary alignment was mainly attributable to its performance in the second cohort. However, this finding should be interpreted cautiously, as it may reflect not only differences in output quality but also changes in students' familiarity with different generative AI platforms over time. The increasing visibility and use of other platforms, such as Gemini, may have influenced the generated outputs.

Table 2: Proportion of AI-generated outputs coded as aligned by researchers across systems and cohorts.

Cohort	*ChatGPT aligned outputs*	*Gemini aligned outputs*	*Meta AI aligned outputs*
First cohort	74.4%	74.4%	65.4%
Second cohort	62.3%	84.2%	71.1%
Total	67.2%	80.2%	68.8%

Gemini was therefore the system most frequently associated with taxonomically coherent outputs in the overall dataset. This pattern is

consistent with students' perceptions of coherence in the second cohort, where it also obtained the highest mean score on the Likert scale. The convergence between students' perceived coherence and the researchers' alignment coding strengthens the interpretation that, within this specific dataset, Gemini generated outputs that were more frequently perceived and evaluated as taxonomically coherent. At the same time, the cohort-level distribution calls for a cautious interpretation, as the relative performance of the systems varied across the two cohorts. This variability suggests that the observed differences should not be interpreted as stable characteristics of the systems, but rather as outcomes situated within the specific conditions of data generation and within the possible evolution of AI systems over time.

On the basis of the binary alignment coding conducted by the researchers, aligned outputs were further analysed according to taxonomic level. As shown in Table 3, the highest levels of alignment were observed for *Understand* and *Remember*, which appeared to represent the most stable cognitive processes across the two cohorts. By contrast, *Create* consistently showed the lowest proportion of aligned outputs, with almost identical values in the first and second cohorts. *Apply* and *Analyse* occupied an intermediate position, although *Analyse* showed a decrease in alignment in the second cohort. *Evaluate*, instead, displayed a different pattern, with a marked increase in aligned outputs in the second cohort.

Table 3: The proportion of AI-generated outputs coded as aligned by the researchers for each taxonomic level, considering the three systems together within each cohort

Taxonomic level	*First cohort*	*Second cohort*	*Total*
Remember	84.6%	86.0%	85.4%
Understand	92.3%	89.5%	90.6%
Apply	66.7%	64.9%	65.6%
Analyse	69.2%	57.9%	62.5%
Evaluate	64.1%	86.0%	77.1%
Create	51.3%	50.9%	51.0%

5. Discussion and conclusion

The results of this study contribute to the debate on the role of generative artificial intelligence in initial teacher education by shifting the focus from the

technical capacity of AI systems to generate instructional materials to their ability to operate consistently within a structured pedagogical framework. From this perspective, the study is not limited to comparing the performance of three AI systems, but rather investigates the extent to which AI-generated learning activities can be meaningfully aligned with the cognitive process dimension of Anderson and Krathwohl's taxonomy.

The findings suggest that cognitive complexity may contribute to shaping the alignment between AI-generated activities and taxonomic levels, although not according to a fully linear pattern. Across the two academic years considered, lower-order processes, such as Remember and Understand, generally showed greater stability and appeared less exposed to ambiguity, which may be related to the more explicit, circumscribed, and easily recognisable nature of the cognitive operations they involve. By contrast, Evaluate showed stronger alignment in the second cohort, whereas Create consistently emerged as the most challenging level, in line with recent literature (Akdeniz et al., 2025; Zhang et al., 2025). Since this pattern emerged across the different systems analysed, the criticalities observed appear to be attributable not only to the specific characteristics of individual models, but also to broader structural constraints of currently available generative AI systems. Tasks labelled as creative often took the form of open-ended productions, such as creating a poster, a product, or an artefact, without necessarily requiring students to reorganise knowledge, integrate conceptual elements, or develop an original structure grounded in disciplinary content. In these cases, the activity appeared formally complex, but did not necessarily entail genuine cognitive complexity.

Similarly, tasks associated with Analyse or Apply were sometimes reduced to comparison, classification, or the execution of familiar procedures, thereby weakening their taxonomic specificity and blurring the distinction between analysis and lower-level cognitive processes. The presence of cases in which none of the three systems produced an adequately aligned output is also noteworthy. These cases should be interpreted with caution, since they may have been influenced by prompt formulation, disciplinary topic, and the inherent variability of AI-generated responses. Nevertheless, their occurrence reinforces the need for teachers to critically examine AI-generated materials before integrating them into instructional design, with particular attention to the cognitive operations actually required by the task.

Finally, one implication of the study lies in the recognition of the pedagogical value of misalignment itself. When subjected to critical analysis, AI-generated

tasks activated processes of discussion, argumentation, and revision, thereby contributing to the development of prospective teachers' pedagogical–disciplinary knowledge and metacognitive awareness. Far from functioning as mere shortcuts for instructional planning, artificial intelligence systems proved to be most effective when integrated into reflective training devices that required students to consciously question, problematise and critically rework the outputs produced.

References

Abu Eideh, B. (2026). Reimagining instructional design: ChatGPT, Bloom's taxonomy, and the role of teacher expertise. *Social Sciences & Humanities Open, 13*, 102325. https://doi.org/10.1016/j.ssaho.2025.102325

Akdeniz, H., Clark, T., & Roberts, J. L. (2025). Can AI generate questions aligned with Bloom's taxonomy? A framework for gifted education to support teachers. *Journal of Advanced Academics, 36*(4), 671–694. https://doi.org/10.1177/1932202X251349917

Almatrafi, O., & Johri, A. (2025). Leveraging generative AI for course learning outcome categorization using Bloom's taxonomy. *Computers and Education: Artificial Intelligence, 8*, 100404. https://doi.org/10.1016/j.caeai.2025.100404

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Educational Evaluation, 31*(2–3), 102–113. https://doi.org/10.1016/j.stueduc.2005.05.004

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Choi, G. W., Kim, S. H., Lee, D., & Moon, J. (2024). Utilizing generative AI for instructional design: Exploring strengths, weaknesses, opportunities, and threats. *TechTrends, 68*(4), 832–844. https://doi.org/10.1007/s11528-024-00967-w

d'Alonzo, L. (2016). *La differenziazione didattica per l'inclusione: Metodi, strategie, attività*. Erickson.

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies

on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31–54. https://doi.org/10.1016/j.edurev.2018.02.002

Farhood, H., Nyden, M., Beheshti, A., & Muller, S. (2025). Artificial intelligence-based personalised learning in education: A systematic literature review. *Discover Artificial Intelligence, 5*(1), 331. https://doi.org/10.1007/s44163-025-00598-x

Karaman, M. R., & Göksu, İ. (2024). Are lesson plans created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education, 7*(1), 107–127. https://doi.org/10.46328/ijte.607

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL IOE Press.

Mendis, M. S., Wijesinghe, C., Madhushara, S., Dilhara, S. R., & Edirisinghe, M. (2025). Redefining pedagogy for the 21st century: A framework combining Bloom's taxonomy, outcome-based education, and active learning strategies. *International Journal of Scientific Engineering and Science, 9*(11), 79–87.

Rose, T. (2016). *The end of average: How to succeed in a world that values sameness*. HarperOne.

Tomlinson, C. A. (2014). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). ASCD.

UNESCO. (2024). *AI competency framework for teachers*. https://doi.org/10.54675/ZJTE2084

van der Graaf, J., Raković, M., Fan, Y., Lim, L., Singh, S., Bannert, M., Gašević, D., & Molenaar, I. (2023). How to design and evaluate personalized scaffolds for self-regulated learning. *Metacognition and Learning, 18*, 783–810. https://doi.org/10.1007/s11409-023-09361-y

Weng, X., Ye, H., Dai, Y., & Ng, O. (2024). Integrating artificial intelligence and computational thinking in educational contexts: A systematic review of

instructional design and student learning outcomes. *Journal of Educational Computing Research, 62*(6), 1420–1450. https://doi.org/10.1177/07356331241248686

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, G., Ying, Y., Jiang, S., Liang, J., Yue, G., Fu, Y., Hu, H., & Xiao, Y. (2025). From remembering to metacognition: Do existing benchmarks accurately evaluate LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 13440–13457). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.724