SAMYRAD 2026

SMART ALGORITHMS & MODELS: YOUR REACH TO ADVANCED DIALOGUES · 5–6 OCTOBER 2026 · SEVILLE, SPAIN
SAMYRAD OPEN-ACCESS PROCEEDINGS | PUBLISHER: INTELLIGENT NUDGE

PAPER • OPEN ACCESS

# Transforming University Teaching through Generative AI: An Exploratory Pilot Study and a Multidisciplinary Scale-Up Design

Marc Simon-Altaba
Faculty of Law, Universitat Internacional de Catalunya (UIC Barcelona), Barcelona, Spain

Article DOI: 10.68286/samyrad2026.327

**ABSTRACT**
Generative artificial intelligence is transforming higher education, but its educational value depends on whether it supports or displaces the reasoning that courses are intended to develop. The objective of this completed exploratory pilot was to examine how a bounded, verification-based integration of GenAI operated in first-year Business Law teaching, focusing on students' perceptions, recurrent limitations in AI-generated outputs, and students' capacity to identify and correct those limitations. The pilot was implemented at UIC Barcelona in two first-year Business Administration groups, one in Spanish and one in English. Case-based activities combined independent work, bounded AI assistance, mandatory verification, and justification of what students accepted, corrected, or rejected. Evidence comprised 65 anonymous end-of-course questionnaires, documented instructor observations, continuous-assessment activities, and an instructor activity log. Students evaluated the experience positively and reported gains in critical AI use but did not perceive an automatic performance advantage and strongly preferred an occasional or regular but supervised role for AI (94.0% and 87.5%). Documented classroom episodes showed three recurrent limitations: generic and uniform outputs in AI-assisted trademark creation, formally plausible but insufficiently specific franchise contracts, and occasional reliance on outdated legislation. These findings support a model of AI as a contestable "third voice": a source of propositions that must be tested against disciplinary criteria rather than treated as an authority or substitute for the student's final intellectual work. The VI Exea Impact Award 2026 has enabled a subsequent multidisciplinary and international scale-up involving UIC Barcelona, Universitat de Barcelona, Universitat de Girona, and United Arab Emirates University. No outcome data from that phase was available at submission. Its purpose is to test the model's transferability through a two-cycle mixed design and a common analytical-reasoning rubric. The present paper therefore reports pilot evidence, and a candidate model whose multidisciplinary and international transferability remains to be tested.

**KEYWORDS** Generative AI; Teaching innovation; Higher education; Case-based learning; Analytical reasoning

**How to cite:** Simon-Altaba, M. (2026). Transforming university teaching through generative AI: An exploratory pilot study and a multidisciplinary scale-up design. SAMYRAD 2026 Open-Access Proceedings. Intelligent Nudge. https://doi.org/10.68286/samyrad2026.327

## Introduction

Generative artificial intelligence (GenAI) has moved rapidly from an optional digital aid to an ordinary component of students' academic environment. The resulting challenge is not simply whether universities should permit its use. It is how teaching and assessment should be designed when students can obtain fluent, plausible, and complete responses within seconds. International guidance has therefore emphasized human agency, institutional validation, and data protection,

while the research literature identifies both opportunities for feedback and engagement and risks involving inaccuracy, academic integrity, dependency and reduced epistemic effort [1]–[3]. This challenge also has an inclusion dimension: access to the technology is insufficient if universities do not develop students' capacity to understand and use it critically [4].

The evidence is also more conditional than early claims of transformation sometimes suggest. A recent analysis shows positive average effects on academic performance, motivation, and higher-order thinking, but also identifies methodological weaknesses and the need to distinguish the quality of an AI-assisted product from traditional learning [5]. A 2026 systematic review similarly concludes that critical and creative thinking are most likely to benefit when GenAI is embedded in inquiry-oriented tasks with reflective prompts, rubrics, and verification requirements; unstructured use is more often associated with cognitive offloading [6]. Student perception studies show broad interest in GenAI but persistent concerns about reliability, privacy, and over-reliance [7]. These findings support a pedagogical rather than technological question: what division of intellectual labour between student, teacher and AI works best for what a course is intended to achieve?

This question is particularly visible in legal and business education. Legal analysis requires students to identify legally relevant facts, determine applicable rules, distinguish authority from assertion, apply criteria to a concrete case, and justify a conclusion. GenAI may generate an answer that reads convincingly while omitting a requirement, inventing a source, overlooking an exception, or failing to adapt general information to the facts. Its fluency can therefore create an illusion of competence precisely where novice students lack the domain knowledge needed to detect the defect. The same problem extends to economics, communication and the humanities whenever learning outcomes depend on interpretation, evaluation of evidence and reasoned judgment rather than textual production alone. Spanish legal-education scholarship similarly treats GenAI as a transformation of both legal learning and professional practice, reinforcing the need to connect technological experimentation with disciplinary judgment [8].

The primary contribution of this paper is the analysis of a completed exploratory pilot conducted at UIC Barcelona in two first-year Business Administration groups taking Business Law. The empirical evidence consists of 65 valid anonymous survey responses, documented instructor observations, an instructor activity log, and evidence arising from ordinary continuous-assessment activities that compared work completed with and without AI. The pilot examines how bounded AI use operated in the classroom and what recurrent limitations emerged when students were required to verify and correct generated material.

The pilot has subsequently informed a larger multidisciplinary and international project whose development received a decisive institutional boost through the VI Beca Exea Impact 2026. The expanded phase connects Law, Economics, Communication and Humanities at UIC Barcelona, Law at Universitat de Barcelona (UB), Economics at Universitat de Girona (UdG), and Economics at United Arab Emirates University (UAEU). This scale-up is presented here as a research implication and future validation stage, not as an empirical component of the present study. No outcome data from the expanded phase was available at the conference deadline, and the paper does not claim that multidisciplinary or international transferability has already been demonstrated.

Three research questions organize the analysis: RQ1 asks how students perceived the effects of structured GenAI activities on interest and understanding; RQ2 identifies recurrent limitations in AI output during case-based work; and RQ3 examines how a supervised model encouraged verification and responsible use. These questions are addressed through complete pilot evidence. The subsequent multidisciplinary phase is presented further in the manuscript in Section ‘From Pilot To Exea-Supported Multidisciplinary Scale-Up’ as the design for testing whether the candidate model transfers across disciplines, institutions, and national contexts.

## Pedagogical and governance framework

### From Output Generation to Reasoning Support

The model begins from constructive alignment: AI use is justified only when the activity still requires the learner to perform the reasoning expressed in the learning outcomes. Case and problem-based approaches are suitable because they organize learning around interpretation, hypothesis formation, application of knowledge and justified decision-making [9], [10]. Formative feedback is educationally valuable when it helps students compare current performance with a standard and regulate their next action, rather than merely supplying a finished response [11]. Evidence from law-school assessment likewise illustrates that apparent task competence does not make the system a dependable source of legal judgment [12]. GenAI can contribute to this process, but only if its role is deliberately limited.

The basic activity sequence is therefore: an independent or minimally assisted first attempt; a controlled AI intervention; verification of the output against course materials, authoritative sources, or an agreed rubric; correction and counterargument; and a final position for which the student remains responsible. Within that sequence, AI may generate questions, challenge an assumption, propose an alternative scenario, audit coherence, or provide a draft for criticism. As we will see, it must not operate as the sole authority, the undisclosed author of the final work or the final assessor of student performance. This division of labour treats GenAI as a contestable cognitive artifact rather than an answer service.

The distinction matters because assistance during a task is not equivalent to learning. Experimental evidence indicates that students can rely on AI rather than internalize the supported process [13]. A large field experiment in mathematics found that unrestricted assistance improved

practice performance but could impair later unaided performance; safeguards that withheld direct solutions and provided teacher-designed hints mitigated the negative effect [14]. Although that setting differs from university legal education, the mechanism is directly relevant: when the tool removes the need to reason, the quality of the immediate product may rise while the learner's capacity remains unchanged or weakens. The present model therefore evaluates not only what students produce with AI, but also how they interrogate it and what they can justify independently.

### Integrity, Human Oversight and Data Protection

A supervised model requires clear ex ante rules. Each task should specify whether AI is prohibited, optional or required; which functions are permitted; what use must be declared; and what evidence of the process must be retained. This is more informative than a binary institution-wide ban because the acceptable degree of AI involvement depends on the learning outcome. An AI assessment scale offers a related approach by linking levels of permitted use to assessment design and transparency [15]. Academic integrity then concerns truthful attribution, preservation of the student's intellectual contribution and accountability for submitted claims, not merely whether a tool was opened [16], [17]. This allocation of responsibility is also consistent with broader community-wide ethical frameworks for AI in education [18].

Human oversight is equally structural. The European Union's AI Act identifies educational uses that may affect access, evaluation or monitoring as especially sensitive and requires safeguards appropriate to the risk; its broader human-oversight logic includes awareness of automation bias [19]. The European Commission's updated 2026 guidelines likewise connect ethical AI use in education with critical AI literacy, the AI Act and the General Data Protection Regulation [20]. In the present model, no AI system makes final grading decisions, and instructors remain responsible for the pedagogical level, legal accuracy and fairness of materials and assessment.

Data protection is implemented through minimization and purpose limitation. Students are instructed not to enter identifiable personal data, confidential material, or sensitive information into external systems. Activities are designed so that fictitious or sufficiently abstracted cases can achieve the learning objective. This approach is consistent with Spanish and EU data-protection guidance emphasizing compliance from the design stage when AI is incorporated into processing operations [21]. Proportionality also has a sustainability dimension: the project favours purposeful, bounded interactions over indiscriminate generation and treats computational use as justified by a defined pedagogical function.

## Completed pilot: method

### Context and Design

The completed phase was an exploratory classroom intervention in an introductory Business Law subject within the first year of the Bachelor's Degree in Business Administration at UIC Barcelona. The intervention was implemented in two parallel groups, one taught in Spanish and one in English. This context was pedagogically demanding, as students were acquiring foundational legal concepts while learning to apply them to business situations, and therefore faced a heightened risk of accepting plausible but incomplete AI explanations. The study combined descriptive survey evidence with documented instructor observation, an instructor activity log, and evidence arising from ordinary continuous-assessment activities. It was not designed as a randomized or quasi-experimental test of causal impact. The groups were defined by language of instruction, not by assignment to an AI treatment, and the available performance comparisons were embedded in ordinary teaching rather than standardized as equivalent outcome measures. The completed pilot is therefore characterized cautiously; the prospective Exea-supported project, rather than this phase, adopts the mixed quasi-experimental design described later.

### Intervention Activities and Safeguards

The first activity type used structured comparison in practical cases. Students alternated between solving selected problems with and without AI under specified conditions. They then compared the solutions and identified omissions, unsupported assumptions, excessive generality, and conceptual errors. The exercise did not reward the mere use of AI. It assessed whether students could determine which parts of an output were legally relevant and explain why a proposed answer should be accepted, corrected, or rejected.

A second activity addressed industrial property. Students used AI to propose elements of a trademark, such as a name or a logo, and then applied legal criteria concerning distinctiveness, descriptiveness and registrability. The educational value lay in separating generation from evaluation: an apparently creative proposal still required analysis of whether it could perform the legal function of a trademark and what obstacles might arise.

A third activity concerned commercial contracting. Students first worked on the structure of a contract, including a franchise-type example, and later compared their work with an AI-generated draft. The comparison revealed a recurrent pattern: AI often produced text with a credible formal appearance but insufficient adaptation to the transaction, missing clauses, weak treatment of risk or generic formulations. Students had to identify these deficiencies and justify revisions rather than submit the generated draft as a finished instrument.

The intervention used several AI and AI-assisted tools according to the pedagogical purpose of each activity. ChatGPT Pro and Perplexity supported the creation and variation of practical cases and assessment questions.

Kahoot Plus supported AI-assisted group assessment games, while Gamma Plus was used to review and improve classroom presentation materials. Various web-based generative-design tools were used in industrial-property activities to create logos, trademarks, and related visual elements. The study did not compare platforms as technological products; each tool was treated as a pedagogical instrument subject to the same verification requirements.

The same toolset also supported the professor in producing preliminary quiz questions, interactive revision items, and variants of case statements. Every item was reviewed for accuracy, level, ambiguity, and alignment with the syllabus. This teacher-facing use accelerated drafting but did not transfer pedagogical judgment to the system; many proposed questions were too easy, overly general, or ambiguous and required substantial calibration.

Three safeguards applied across activities. First, students had to verify all AI outputs against course materials and subjects explained in class. Second, the final answer required a reasoned account of what was retained, changed, or discarded. Third, some tasks restricted AI use so that students preserved opportunities for independent performance and could compare different working processes. No automated grading was used, and the AI output was never treated as an authoritative legal source.

### Participants, Data and Analysis

An anonymous end-of-course questionnaire produced 65 valid responses: 33 from the Spanish-language group and 32 from the English-language group. The questionnaire addressed prior frequency of AI use, overall evaluation of the experience, perceived effects on interest and understanding, comparison of practical work with and without AI, development of critical-use habits, detection of errors or incomplete responses, and the preferred role of AI in teaching.

During implementation, practical activities formed part of the ordinary continuous-assessment process. The professor considered not only the substantive final answer but also students' capacity to detect gaps, omissions, excessive generality, or unsupported reasoning in AI-generated material and to justify the corrections they proposed. In parallel, the instructor maintained an activity log recording the exercises implemented, recurrent characteristics of AI-generated outputs, classroom comparisons, and the principal difficulties observed. These records were used to document recurring patterns and select illustrative classroom episodes; they were not treated as an independently coded qualitative dataset.

Analysis was descriptive. The questionnaire was administered through Moodle, whose reporting functionality automatically generated the response counts and percentages used in the analysis; no separate statistical package was used. Percentages were examined within each language group and are reported without inferential comparison. Because the questionnaire measured perceptions and self-reported learning, it cannot establish objective gains in analytical reasoning. Instructor observations, the activity log, and continuous-assessment evidence were used descriptively to identify recurring patterns and illustrative episodes, but no independent coding or inter-rater reliability procedure was applied. No directly identifying information was collected for the analysis, and students were instructed not to include personal or confidential information in external AI tools.

## Results

### Survey and Descriptive Findings

The first finding is that AI was already normalized in students' academic practice. In the Spanish-language group, 48.5% reported using it fairly frequently and 33.3% habitually; the corresponding figures in the English-language group were 34.4% and 43.8% (Table 1). Thus, 81.8% and 78.2% respectively fell within the two highest frequency categories. The pedagogical problem was not how to introduce an unfamiliar technology, but how to convert existing use into a more accountable and evaluative practice.

**Table 1.** Selected student survey findings

| Indicator | Spanish Participants=33 | English Participants=32 |
|---|---|---|
| Prior use: fairly frequent or habitual | 81.8% | 78.2% |
| Experience: positive or very positive | 75.8% | 81.3% |
| Course more interesting: "quite" | 60.6% | 50.0% |
| Understanding improved: quite or much | 78.8% | 68.8% |
| AI work: results "similar" | 48.5% | — |
| AI work: "depends on activity" | — | 46.9% |
| Critical AI use learned: quite or much | 72.7% | 71.9% |
| Errors found: quite often or most times | 81.8% | 56.3% |
| Prefer occasional or regular supervised role | 94.0% | 87.5% |

The overall evaluation was favourable. In the Spanish group, 69.7% described the experience as positive and 6.1% as very positive; in the English group, the proportions were 59.4% and 21.9%. Students also reported motivational and comprehension benefits. The course was considered "quite" more interesting by 60.6% of the Spanish group and 50.0% of the English group. For perceived understanding, 63.6% of the Spanish group selected "quite" and 15.2% "much"; in the English group, 43.8% selected "quite" and 25.0% "much." These are perception measures, but they answer RQ1 by indicating that the structured activities were broadly

accepted and associated with greater interest and perceived comprehension.

The data do not support an automatic performance advantage. Among the Spanish-group participants in the survey, 48.5% considered the results of AI-supported practical work similar and 24.2% considered them worse. In the English group, the most frequent answer was that the outcome depended on the activity (46.9%). Instructor observation was consistent with this qualification: AI was more useful for producing a structure or draft than for applying legal criteria accurately to a specific set of facts. Where conceptual knowledge was weak, access to a fluent draft did not resolve the underlying difficulty.

The strongest result concerned critical use. In the Spanish group, 54.5% stated that they had learned to use AI critically "quite" a lot and 18.2% "much"; in the English group, the figures were 56.3% and 15.6%. Students also encountered errors or incompleteness frequently. In the Spanish group, 63.6% reported this on "quite a few" occasions and 18.2% on most occasions. In the English group, the corresponding proportions were 37.5% and 18.8%. These responses support RQ2 and RQ3: recurrent defects were visible to students, and a substantial majority perceived development in their capacity to question the output.

Students' preferred institutional model was correspondingly cautious. In the Spanish group, 48.5% favoured occasional support and 45.5% regular but supervised use. In the English group, 37.5% chose occasional support and 50.0% regular supervised use. A central role for AI was selected by only 6.1% and 9.4%, respectively. The preference is not for exclusion, but for integration within boundaries that preserve teacher oversight and student responsibility.

Across the practical activities, the instructor activity log recorded recurrent limitations including answers that were correct in the abstract but insufficient for the case; omission of legally relevant requirements; imprecise terminology; failure to distinguish closely related concepts; inadequate adaptation of contract language; and confident statements lacking qualification or current authority. These patterns explain why the tool's most productive classroom role was not to supply the correct answer, but to create an object that students could test against disciplinary criteria and official legal sources.

### Illustrative Episodes of AI Output and Student Verification

The activity log and continuous-assessment discussions provide concrete examples of how this verification process operated. These episodes are illustrative rather than an independently coded qualitative dataset, but they show the mechanism underlying the broader patterns reported above.

In the industrial-property exercise, students were asked to create commercial brands for businesses specified by the professor using AI-assisted design tools. The generated proposals were technically usable but generally modest: many were uniform, generic, and insufficiently differentiated from one another. During the subsequent class comparison, students themselves identified this pattern. The exercise therefore shifted from asking whether AI could generate a brand to asking whether the proposal displayed sufficient distinctiveness and contextual fit to function persuasively as a commercial identity.

In the franchise-contract activity, AI-assisted drafts were again formally plausible and structurally correct, but their clauses remained general and lacked depth in relation to the concrete transaction. During classroom comparison, students identified that important provisions required greater specificity, including clauses dealing with indemnification in the event of breach. The problem was not the absence of a contract-like document, but the limited capacity of a generic draft to translate formal structure into sufficiently concrete rights, obligations, and consequences. Students' ability to identify such gaps and justify corrections formed part of the continuous-assessment process.

A third episode concerned temporal legal accuracy. In some practical exercises, classroom review detected that AI-generated answers relied on superseded versions of the applicable legislation. Verification of current course materials and official legal sources was therefore necessary before the reasoning could be accepted. This problem is particularly significant in legal education because fluent explanation can conceal obsolete authority, and it directly supports the model's rule that AI output must never be treated as a legal source in itself. Taken together, the three episodes illustrate different forms of contestability: students had to differentiate a generated proposal, test the substantive adequacy of an apparently complete document, and verify the currency of the authority supporting an answer. Their value lies not in demonstrating causal improvement in critical thinking, which the design cannot establish, but in showing how the classroom tasks made verification and correction observable components of student work.

## Discussion

### AI as a Contestable "Third Voice"

The completed pilot supports a model in which AI functions as a "third voice" in the classroom: neither the teacher, nor the student, nor an authoritative source, but an additional position that must earn acceptance through verification. Its educational value arises from contestability. A draft that is incomplete, overconfident, generic, or outdated becomes useful when the task requires students to locate the defect, reconstruct the reasoning, and defend a correction. The same output becomes educationally weak when it is copied because it appears plausible. The trademark, franchise-contract, and outdated-legislation episodes illustrate three forms of contestability: differentiating a generated proposal, testing the substantive adequacy of an apparently complete document, and verifying the temporal validity of the authority supporting an answer.

This interpretation reconciles the favourable student experience with the absence of a perceived automatic performance advantage. Motivation and perceived understanding may improve because the tool introduces variety, immediacy, and alternative formulations. Yet better assisted production does not itself demonstrate stronger unaided reasoning. That distinction is central in recent evidence reviews [5], [6] and in methodological critiques of studies that do not define the treatment, control condition, and learning outcome precisely [22]. The present pilot therefore treats the neutral or task-dependent performance result as substantively informative, not as a failure of innovation.

### Guardrails Preserve the Learning Task

The verification and justification rules appear to be the mechanism most consistent with the reported development of critical habits. They prevent the students' contribution from collapsing into prompt formulation and blind acceptance of outputs. This aligns with evidence that unrestricted assistance can become a cognitive crutch, whereas carefully designed constraints can preserve engagement [14]. It also responds to systematic concerns that over-reliance may weaken decision-making and analytical judgment [23]. The relevant unit of design is thus not "use of ChatGPT" in the abstract, but a complete instructional sequence specifying what the student must do before, during and after consultation.

For assessment, this implies that authenticity should be evidenced through process as well as product. Suitable measures include a short AI-use declaration, retained prompts or versions where proportionate, an annotated verification note, identification of rejected suggestions, oral defence of a key decision and an unaided component. These measures should not become surveillance by default; they should be selected only where they validly reveal the learning outcome. Long-term assessment design should connect present tasks with the independent judgment that graduates will need beyond their studies [24].

### The Professor Remains Central

The pilot also shows that GenAI changes rather than eliminates academic work. The instructor spends less time producing a first draft of a quiz or scenario, but more time calibrating difficulty, checking accuracy, anticipating misuse, and designing the comparison that will make the output pedagogically productive. Therefore, expertise from professors remains indispensable because only a subject specialist can judge whether a coherent response captures the doctrinal requirements, evidential standards, and level appropriate to the course.

This shift has institutional implications. Developers should focus not only on operating tools, but on task architecture, verification protocols, data protection, and assessment validity. Policy should allow differentiated uses rather than assuming that one permission rule fits every subject or learning outcome [17]. Clear communication also reduces inequity: students should not have to infer from informal practice whether AI is acceptable, which functions are allowed or how use must be acknowledged.

### Transferability and Disciplinary Difference

The pilot's underlying reasoning cycle is potentially transferable because it is not tied to a single legal doctrine: identify relevant information; distinguish evidence from assumption; apply a disciplinary standard; test alternatives; justify a conclusion; and state uncertainty. However, transferability does not mean uniformity. Law emphasizes authority, qualification, and application to facts; economics may emphasize model assumptions, causal interpretation, and sensitivity to data; communication may emphasize audience, framing and source credibility; and the humanities may emphasize interpretation, contextualization, and competing readings. A future common rubric must therefore capture general reasoning dimensions while allowing discipline-specific indicators and examples. This is also an employability issue: technological systems are expected to restructure professional work, making expert judgment, verification, and accountability more rather than less important [25].

On top of that, the international extension adds further variables: language, local regulation, educational conventions, and differences in students' prior exposure to GenAI. For instance, UAEU's participation can test whether the model operates across a different institutional and national setting, but it also requires equivalence procedures for instruments and careful interpretation of contextual effects. The completed pilot justifies testing the model, yet it does not establish that it will transfer unchanged.

## From pilot to Exea-supported multidisciplinary scale-up

The completed pilot subsequently became the basis for a broader research and teaching-innovation program. The VI Beca Exea Impact 2026 marked an inflection point by supporting the development of common instruments and a network extending beyond the original Business Law classroom. The current scope includes Law, Economics, Communication and Humanities at UIC Barcelona; Law at UB; Economics at UdG; and Economics at UAEU. This configuration creates three forms of transfer to be tested: across disciplines within one institution, across universities within Spain, and across national and educational contexts through UAEU.

The next phase is designed to test rather than assume transferability. Over two academic years, a first formative cycle will adapt cases, prompts, declarations and rubric descriptors to selected subjects and refine implementation procedures; a second cycle will use the revised materials to generate more comparable evidence. The shared core will preserve bounded AI roles, explicit permitted uses, verification, and a final accountable student judgment. The principal planned outcome is an analytical reasoning index based on argumentative structure, integration of evidence,

treatment of counterarguments, justificatory coherence, and explicit handling of uncertainty, with discipline-specific descriptors where necessary.

The mixed evaluation will combine objective performance tasks with pre- and post-questionnaires, declared AI-use records, student focus groups, staff interviews and, where feasible and proportionate, aggregated learning-platform data. Selected unaided components are intended to distinguish assisted output from retained capability, while qualitative evidence will examine how students decide whether an AI output is trustworthy and how teachers calibrate the tool's role. Any use of learning-platform data will be governed by proportionality, privacy, transparency and ethical review [26]. No outcome data from this expanded phase is available in the present paper. The pilot has therefore generated a structured candidate model and testable hypotheses for multidisciplinary and international evaluation, not evidence that such transfer has already succeeded.

## Limitations

The completed study has several limitations. The sample was limited to 65 respondents from two groups in one first-year subject at one institution. Participation reflected ordinary course conditions, and the questionnaire captured perceptions rather than a validated measure of learning. The language groups were not randomized treatment groups and were not compared inferentially. The activity log provides documented instructor observations, but these were not independently coded or subjected to inter-rater analysis. Continuous-assessment activities were not designed as standardized outcome measures and are used here to illustrate classroom mechanisms, not to demonstrate gains. The study did not include standardized pre- and post-tests, a control condition, or independently assessed performance measures. Consequently, causal claims about motivation, understanding, critical thinking, or analytical reasoning would be unwarranted.

The intervention also combined several activities and uses of AI, so the descriptive findings cannot isolate which component produced a particular perception or classroom response. Tool behaviour may change with model versions, language, and prompts, limiting technological reproducibility. The three classroom episodes reported in the Results Section are documented illustrations rather than a formally coded qualitative sample. Finally, the expanded project is prospective. Its broader institutional composition increases relevance but also introduces disciplinary, linguistic, institutional, and national variability that the common protocol must manage. These limitations explain the scale-up's emphasis on defined treatments, comparable measures, objective reasoning outcomes, and two cycles of refinement.

## Conclusion

The pilot indicates that GenAI can contribute to university teaching without becoming a substitute for student reasoning, but the benefit is conditional on design. Students evaluated the experience positively and reported greater interest, perceived understanding, and critical-use habits. At the same time, they did not perceive an automatic advantage in practical results and overwhelmingly preferred an auxiliary, supervised role for AI. The documented classroom episodes add practical detail to those perceptions: AI-assisted brands tended toward uniformity and generality, franchise-contract drafts were formally plausible but insufficiently specific, and some practical answers relied on outdated legislation. The most defensible interpretation is therefore not that GenAI produces better learning by itself, but that it can create productive occasions for verification, correction, and justified decision-making.

The contestable "third voice" model captures this contribution. AI proposes; the student evaluates; pre-established disciplinary sources and criteria determine whether the proposal is adequate; and the professor designs, documents, and supervises the process. This architecture preserves human accountability and makes the imperfections of the tool part of the learning task rather than a hidden risk. The current pilot provides descriptive evidence of how that mechanism operated, while its limitations preclude claims of causal improvement in analytical reasoning.

The VI Exea Impact Award 2026 has accelerated the project's evolution from a local legal-education pilot to a multidisciplinary and international evaluation program involving UIC Barcelona, UB, UdG and UAEU. The next phase will test whether the candidate reasoning framework can transfer across Law, Economics, Communication and Humanities and across different institutional and national contexts. Its contribution will depend on maintaining the evidential discipline applied here: separating assisted performance from learning, completed results from expected outcomes, and technological novelty from the pedagogical conditions that make it educationally useful.

## Acknowledgment

The author acknowledges the institutional support associated with the VI Beca Exea Impact 2026 and the contribution of the participating teaching teams at UIC Barcelona, Universitat de Barcelona, Universitat de Girona and United Arab Emirates University to the development of the expanded project.

## References

[1] UNESCO, Guidance for Generative AI in Education and Research. Paris, France: UNESCO, 2023.

[2] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, Art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[3] M. Giannakos et al., "The promise and challenges of generative AI in education," Behav. Inf. Technol., vol. 44, no. 11, pp. 2518–2544, 2025, doi: 10.1080/0144929X.2024.2394886.

[4] P. Gallego Rodríguez, "El derecho constitucional a la educación y la IA generativa: propuestas para mitigar la exclusión digital educativa," Estudios de Deusto, vol. 73, no. 1, pp. 183–212, 2025, doi: 10.18543/ed.3329.

[5] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu, "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies," Comput. Educ., vol. 227, Art. no. 105224, 2025, doi: 10.1016/j.compedu.2024.105224.
[6] C. Li, H. Cui, and L. S. Hagedorn, "The cognitive impact of ChatGPT in higher education: A systematic review of critical and creative thinking outcomes," Comput. Educ.: Artif. Intell., vol. 10, Art. no. 100571, 2026, doi: 10.1016/j.caeai.2026.100571.
[7] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: Perceptions, benefits, and challenges in higher education," Int. J. Educ. Technol. Higher Educ., vol. 20, Art. no. 43, 2023, doi: 10.1186/s41239-023-00411-8.
[8] A. Abadías Selma and D. González Uriel, Eds., El impacto de la IA en el aprendizaje y en la práctica del Derecho. Las Rozas, Spain: Aranzadi LA LEY, 2024.
[9] C. E. Hmelo-Silver, "Problem-based learning: What and how do students learn?" Educ. Psychol. Rev., vol. 16, pp. 235–266, 2004, doi: 10.1023/B:EDPR.0000034022.16470.f3.
[10] J. Biggs and C. Tang, Teaching for Quality Learning at University, 4th ed. Maidenhead, U.K.: Open Univ. Press, 2011.
[11] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," Stud. Higher Educ., vol. 31, no. 2, pp. 199–218, 2006, doi: 10.1080/03075070600572090.
[12] J. H. Choi, K. E. Hickman, A. B. Monahan, and D. Schwarcz, "ChatGPT goes to law school," J. Legal Educ., vol. 71, no. 3, pp. 387–400, 2022.
[13] A. Darvishi, H. Khosravi, S. Sadiq, D. Gašević, and G. Siemens, "Impact of AI assistance on student agency," Comput. Educ., vol. 210, Art. no. 104967, 2024, doi: 10.1016/j.compedu.2023.104967.
[14] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics," Proc. Natl. Acad. Sci. U.S.A., vol. 122, no. 26, Art. no. e2422633122, 2025, doi: 10.1073/pnas.2422633122.
[15] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, "The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment," J. Univ. Teach. Learn. Pract., vol. 21, no. 6, 2024, doi: 10.53761/q3azde36.
[16] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," Innov. Educ. Teach. Int., vol. 61, no. 2, pp. 228–239, 2024, doi: 10.1080/14703297.2023.2190148.
[17] C. K. Y. Chan, "A comprehensive AI policy education framework for university teaching and learning," Int. J. Educ. Technol. Higher Educ., vol. 20, Art. no. 38, 2023, doi: 10.1186/s41239-023-00408-3.
[18] W. Holmes et al., "Ethics of AI in education: Towards a community-wide framework," Int. J. Artif. Intell. Educ., vol. 32, pp. 504–526, 2022, doi: 10.1007/s40593-021-00239-1.
[19] European Parliament and Council, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Off. J. Eur. Union, L 2024/1689, Jul. 12, 2024.
[20] European Commission, Guidelines on the Ethical Use of Artificial Intelligence and Data in Teaching and Learning. Brussels, Belgium: European Commission, 2026.
[21] Agencia Española de Protección de Datos, Una aproximación para la adecuación al RGPD de tratamientos que incorporan inteligencia artificial. Madrid, Spain: AEPD, Feb. 2020.
[22] J. Weidlich, D. Gašević, H. Drachsler, and P. Kirschner, "ChatGPT in education: An effect in search of a cause," J. Comput. Assist. Learn., vol. 41, no. 5, Art. no. e70105, 2025, doi: 10.1111/jcal.70105.
[23] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review," Smart Learn. Environ., vol. 11, Art. no. 28, 2024, doi: 10.1186/s40561-024-00316-7.
[24] D. Boud and N. Falchikov, "Aligning assessment with long-term learning," Assess. Eval. Higher Educ., vol. 31, no. 4, pp. 399–413, 2006, doi: 10.1080/02602930600679050.
[25] R. Susskind and D. Susskind, The Future of the Professions: How Technology Will Transform the Work of Human Experts. Oxford, U.K.: Oxford Univ. Press, 2015, doi: 10.1093/oso/9780198713395.001.0001.
[26] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," Am. Behav. Sci., vol. 57, no. 10, pp. 1510–1529, 2013, doi: 10.1177/0002764213479366.