Does AI Ideation Feel Collaborative? Collaboration Potential as a Structurally Distinct Evaluation Dimension in an LLM-Based Ideation Tool

Stefania Zourlidou, University of Koblenz, Germany
Kamyab Farokhi, University of Koblenz, Germany
Frank Hopfgartner, University of Koblenz, Germany

The European Conference on Education 2026
Official Conference Proceedings

Abstract

Large language models are increasingly used to help students generate project ideas, but an idea that is easy to generate is not necessarily one that students see as useful for working with others. This paper re-analyses a pilot dataset from EduIDEAtor, an LLM-based ideation tool, to ask a narrow question that was not addressed in two earlier papers from the same deployment: does perceived collaboration potential move together with other quality ratings, or does it show a different response pattern? Sixteen students completed the post-use survey, with 14 valid responses for collaboration potential. Collaboration was rated positively: 78.6% selected Better or Much Better than their usual brainstorming method (median = 4 on the intended five-point scale). However, collaboration potential was weakly related to comparative engagement (rho = .27), creativity (rho = -.02), and ease of use (rho = .20). Its only clear association within the comparative block was with project relevance (rho = .65, p = .013). It was not significantly associated with any of the seven broader experience items. Reliability of the five-item comparative block increased from alpha = .814 to .860 when collaboration was removed. With this small sample and a single collaboration item, the result does not establish a separate construct. It does suggest that perceived collaboration potential may capture something different from general usability and engagement, and should be tested directly in future group-based studies.

Keywords: large language models, collaboration, data science education, ideation, project-based learning

iafor
The International Academic Forum
www.iafor.org

Introduction

Large language models are increasingly discussed as tools for teaching and learning. They can support tasks such as generating educational content, answering questions, and providing tutoring support, but they also create risks around inaccurate information and misuse. Kasneci et al. (2023) discuss opportunities, challenges, risks, and mitigation strategies for LLMs in education. Lo (2023) similarly describes uses such as instructor assistance and virtual tutoring, while noting the risk of incorrect or fabricated information. Both reviews discuss learner- and instructor-facing uses of LLMs. The present paper turns to a narrower issue: how students judge the potential use of AI-generated ideas with other people. Project-based learning is designed to engage students in the investigation of authentic problems (Blumenfeld et al., 1991). In collaborative learning, however, putting learners in the same situation does not guarantee that productive collaboration will occur. Dillenbourg (1999) argues that collaborative situations are expected to trigger particular forms of interaction, but these interactions may or may not actually happen. For an ideation tool, this means that generating a useful idea for one student is not the same as supporting discussion and joint development of that idea. Creativity-support research makes a similar distinction. Shneiderman (2007) describes creativity support tools as useful for individual expression, group collaboration, and social creativity. Frich et al. (2019) provide a broad review of 143 HCI papers on creativity support tools and map key characteristics of this design space. Combined, these results suggest that creativity support is not limited to a single mode of use. For an LLM ideation tool, this raises the practical question of whether a student may like the interface and the ideas it produces while still judging those ideas differently when considering their use in a group setting.

Novel Contributions Relative to Earlier Publications

This paper uses the same EduIDEAtor pilot dataset (Farokhi & Zourlidou, 2026) as two earlier papers, but it does not repeat their main analyses. Farokhi et al. (2026) reported the broad pilot evaluation: usability and experience ratings, comparisons with familiar brainstorming, and associations with satisfaction and willingness to reuse the tool. That paper already showed that collaboration potential received positive ratings in the comparative block. A second paper from the same deployment focused on open-ended improvement comments and translated them into interface and implementation priorities, such as recoverability, clearer output presentation, transparency, and more steerable literature suggestions (Zourlidou et al., 2026). The present paper focuses on a more specific question. It does not examine whether students liked EduIDEAtor, whether they would use it again, or which interface improvements should be prioritised. Instead, it concentrates on a narrower aspect of their experience: does perceived collaboration potential rise and fall with engagement, creativity, ease of use, and project relevance, or does it behave differently? The earlier pilot paper reported the overall reliability of the five comparative items, but it did not examine their inter-item structure or what happens to reliability when the collaboration item is removed. The design-oriented paper did not address this question either.The present paper therefore examines whether perceived collaboration potential moves in line with the other comparative quality ratings or instead shows a relatively distinct response pattern within this pilot dataset.

Method

The analysis uses 16 anonymised post-use survey responses collected after students interacted with EduIDEAtor at a German university. Full recruitment, procedure, and tool details are reported in Farokhi et al. (2026), and the anonymised dataset is publicly available (Farokhi &

Zourlidou, 2026). In brief, EduIDEAtor supported early project ideation by taking a student's interests and constraints and returning a small set of candidate directions with short explanations and suggested first steps. The evaluated version was used in a single session before the survey was completed. The survey contained five comparative questions asking students to rate EduIDEAtor against their usual brainstorming method on engagement, creativity, ease of use, project relevance, and collaboration potential. The response anchors were Much Worse, Worse, Same, Better, and Much Better. For this analysis, the verbal responses were coded directly using the released codebook: 1 = Much Worse, 2 = Worse, 3 = Same, 4 = Better, and 5 = Much Better. Coding from the original labels avoids ambiguity about the neutral midpoint and preserves the ordinal meaning of the questionnaire. One respondent left the complete comparative block blank. A second response recorded collaboration potential as the invalid combined value “Same;Better”; this entry was treated as missing for analyses involving collaboration. Therefore, collaboration analyses use $n = 14$. The survey also contained seven five-point experience items covering navigation, input clarity, satisfaction, engagement, match to interests and goals, creativity support, and motivation.

The analysis was deliberately small and descriptive. Spearman rank correlations were used because the survey responses are ordinal and the sample is small. First, collaboration potential was correlated with the four other comparative dimensions. Second, it was correlated with the seven broader experience items. To show the contrast, we also inspected the six pairwise correlations among engagement, creativity, ease of use, and project relevance. Finally, Cronbach's alpha was calculated for the complete five-item comparative block and again after removing collaboration potential. Reliability calculations use the 14 respondents with complete data on all five comparative items. The p-values are reported as descriptive aids rather than as confirmatory evidence. The sample is too small for strong inference, and several related correlations are inspected. The interpretation therefore focuses on the size and pattern of associations, not on a pass/fail significance rule.

Results

Collaboration Ratings

As already suggested by the broad pilot evaluation, perceived collaboration potential was not a weakly rated item. Among the 14 valid responses, the median was 4.0 (IQR = 0.75), and 11 students (78.6%) selected Better or Much Better than their usual brainstorming method. The new result is not this positive average. It is the way the collaboration ratings relate to the rest of the survey. Collaboration potential had weak associations with comparative engagement ($\rho = .27$, $p = .357$), creativity ($\rho = -.02$, $p = .956$), and ease of use ($\rho = .20$, $p = .492$). Project relevance was the exception: it showed a moderate-to-strong positive association with collaboration potential ($\rho = .65$, $p = .013$). The seven broader experience items showed the same general picture. None had a statistically significant association with collaboration potential in this small sample, although satisfaction ($\rho = .47$) and motivation ($\rho = .44$) showed the largest positive coefficients.

Table 1
Spearman Associations With Perceived Collaboration Potential

Variable	n	rho	p
Comparative: Engagement	14	.267	.357
Comparative: Creativity	14	-.016	.956
Comparative: Ease of use	14	.201	.492
Comparative: Project relevance	14	.646	.013
Experience: Easy to navigate	14	.169	.563
Experience: Input was clear	14	.166	.570
Experience: Satisfaction	14	.469	.090
Experience: More engaging than traditional	14	-.140	.633
Experience: Ideas matched interests/goals	14	.348	.223
Experience: Creativity support	14	.337	.239
Experience: Motivation	14	.439	.117

Note. Pairwise available cases are used. The p-values are descriptive and are not adjusted for multiple comparisons.

Non-collaboration Ratings

The four non-collaboration comparative dimensions showed a clearer shared pattern. Their six pairwise Spearman correlations ranged from rho = .47 to .80. Five of the six were statistically significant at $p < .05$: engagement-creativity (.80), engagement-ease of use (.79), creativity-ease of use (.73), creativity-project relevance (.54), and ease of use-project relevance (.58). The remaining association, engagement-project relevance, was positive but weaker (rho = .47, $p = .088$). In other words, the non-collaboration ratings generally rose and fell together, while collaboration potential joined that pattern only through project relevance.

Reliability Check

The reliability check points in the same direction, but should be interpreted cautiously. The five comparative items had Cronbach's alpha = .814. When collaboration potential was removed, alpha increased to .860. Removing an item does not prove that the item measures a separate construct, especially with only 14 complete cases. Here it is simply another indication that collaboration potential adds a different response pattern to an otherwise more internally consistent group of ratings.

Table 2
Reliability of the Comparative Evaluation Block

Scale	Items	Cronbach's alpha
Full comparative block	Engagement, creativity, ease of use, project relevance, collaboration potential	.814
Without collaboration potential	Engagement, creativity, ease of use, project relevance	.860

Note. Both estimates use the same 14 complete comparative responses.

Discussion

The main result is that students could rate collaboration potential positively without rating it in the same way as ease of use, creativity, or engagement. Those other quality judgments mostly moved together. Collaboration did not follow that pattern as consistently. Its strongest and only statistically clear link within the comparative block was project relevance. One possible interpretation is that students use different information when they judge whether an AI-generated idea could work in a group. Interface quality may matter for the experience of generating an idea, while project relevance may matter more when deciding whether the idea is worth taking to peers. This interpretation is compatible with Dillenbourg's (1999) point that collaboration depends on interactions that actually occur, not simply on placing learners in a collaborative situation. It is also consistent with Shneiderman's (2007) distinction between individual expression, group collaboration, and social creativity. This interpretation nevertheless must remain tentative. The survey did not ask students to explain what they meant by collaboration potential, and it did not observe any actual group interaction. The item may reflect imagined shareability, perceived usefulness for discussion, or something else. The data show a response pattern, but they do not tell us the psychological reason for that pattern.

The result suggests a useful design hypothesis. Improving navigation, wording, or visual polish may improve the individual experience without necessarily changing perceived collaboration potential. If future versions of EduIDEAtor aim to support group work, collaboration should be designed and evaluated directly rather than treated as a side effect of good usability. Possible features include a simple way to share an idea card, compare two students' candidate ideas side by side, add short peer comments, or export selected ideas as discussion prompts. These are proposals for future testing, not features shown by the present study to be effective. Their value would need to be assessed through actual collaborative tasks. This is also where the present paper differs from the earlier design-oriented analysis. Zourlidou et al. (2026) used student comments to identify general interface priorities such as recoverability, clearer presentation, transparency, and steerable literature suggestions. The current analysis does not replace those priorities. It adds a separate question for the next design cycle: if collaboration matters, what specific interaction makes an AI-generated idea easier to discuss, compare, revise, or adopt with others? A similar point applies to classroom use. An individual AI ideation step can be followed by a deliberately social step: for example, students can bring one or two candidate ideas to a short peer-review activity and explain why the idea is relevant, feasible, and worth developing. This kind of activity would make collaboration observable rather than assumed. It would also allow future studies to compare perceived collaboration potential with actual behaviours such as discussion quality, revision of ideas, agreement on a shared direction, or division of tasks.

Limitations

Several limitations are important. First, this is a small pilot from one institution. There are 16 survey responses and only 14 valid collaboration responses, so individual answers have a large effect on the estimates. Second, collaboration potential is represented by one comparative item. A single item cannot establish a reliable multi-item construct, and the increase in Cronbach's alpha after removing it does not prove structural independence. Third, the study used a single post-use session and self-report data. Students did not complete a controlled collaborative task, so the paper cannot make claims about actual teamwork, learning gains, coordination, negotiation, or project quality. Fourth, several correlations were inspected in a small dataset. The p-values should therefore be read as descriptive rather than confirmatory. Fifth, the same dataset has been used in earlier publications. We address this by keeping the present contribution narrow and by explicitly separating it from the earlier broad evaluation and design-prioritisation analyses. The new contribution is the the pattern of associations involving collaboration potential, not a new claim that the tool was generally well received.

Conclusion

Perceived collaboration potential received positive ratings in this EduIDEAtor pilot, but it did not behave like most of the other quality ratings. Engagement, creativity, ease of use, and project relevance generally moved together. Collaboration potential had weak or near-zero associations with engagement, creativity, and ease of use, and a clearer association only with project relevance. Removing the collaboration item also increased the internal consistency of the remaining comparative block. With 14 valid responses and a single collaboration item, this should be treated as an exploratory signal rather than evidence for a validated separate construct. The practical value of the result is that it points to a question that broad usability scores can hide. A tool can feel easy and creative without telling us whether its outputs work well in a group. Future studies should therefore measure collaboration directly, using group tasks and collaboration-specific features, and compare perceived shareability with what students actually do together.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

OpenAI's ChatGPT was used during revision to improve language, readability, and organisation. All numerical analyses, interpretations, references, and final wording were checked and approved by the authors. Generative AI was not used to create or modify the study data.

References

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26*(3-4), 369–398. https://doi.org/10.1080/00461520.1991.9653139

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 1–19). Elsevier.

Farokhi, K., & Zourlidou, S. (2026). EduIDEAtor survey: Perceptions of an LLM-based idea development tool [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18171000

Farokhi, K., Zourlidou, S., & Hopfgartner, F. (2026). Does a minimal LLM assistant help students start? A pilot evaluation of an LLM-based ideation tool. In *INTED2026 Proceedings* (Article 1134). IATED. https://doi.org/10.21125/inted.2026.1134

Frich, J., MacDonald Vermeulen, L., Remy, C., Biskjaer, M. M., & Dalsgaard, P. (2019). Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 389, pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300619

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., Stadler, M., Weller, J., Kuhn, J., & 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

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Shneiderman, B. (2007). Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM, 50*(12), 20–32. https://doi.org/10.1145/1323688.1323689

Zourlidou, S., Farokhi, K., & Hopfgartner, F. (2026). Improving an LLM project ideation tool: evidence from a pilot study. In L. Gómez Chova, C. González Martínez, & J. Lees (series eds.), *EDULEARN26 Proceedings: 18th International Conference on Education and New Learning Technologies, 29 June-1 July, 2026, Palma, Spain* (p. 462). IATED Academy. https://doi.org/10.21125/edulearn.2026.0462

Contact email: zourlidou@uni-koblenz.de