Proceedings of the Design Society, Volume 6: DESIGN 2026
https://doi.org/10.1017/pds.2026.10409

Evaluating TRIZ with and without LLM support: an experimental study on engineering problem-solving

Vanja Čok[1], Damien Motte[2], Khadija Hmina[3], Ibtissam El Hassani[3,2,4], Ivan Demšar[1], Jože Tavčar[2] and Nikola Vukašinović[1,✉]

[1] Faculty of Mechanical Engineering, University of Ljubljana, Slovenia, [2] Department of Design Sciences, Lund University, Sweden, [3] ENSAM, University of Moulay Ismail, Morocco, [4] Mathematics, Computer Science and Engineering Department, University of Quebec at Rimouski, Canada

✉ nikola.vukasinovic@fs.uni-lj.si

ABSTRACT: This paper examines integrating Large Language Models (LLMs) into the TRIZ contradiction matrix (TRIZ-C+LLM) to support engineering students in creative problem-solving. Experiments with three problems show that LLMs did not always improve design quality for complex tasks but reduced cognitive workload, improved understanding of contradictions, and increased perceived usefulness. Prompting experience strongly influenced outcomes, highlighting both the promise and limits of combining TRIZ with generative AI.

KEYWORDS: large language model (LLM), artificial intelligence (AI), conceptual design, TRIZ, problem solving

1. Introduction

Creative problem-solving skill is a fundamental competency for driving innovation in competitive markets. Engineering problems can often be modelled as a set of engineering contradictions: improving some design parameters leads to worsening of other design parameters. The contradiction matrix, one of the numerous tools from the Theory of Inventive Problem Solving (TRIZ), developed by Altshuller, offers a structured approach to innovation by resolving contradictions through inventive principles (Orloff, 2020). The TRIZ contradiction matrix connects each contradiction with inventive principles, that is, principles that have been empirically shown to help solve these contradictions. These principles then help the engineers in finding solutions to the original problem.

The use of the TRIZ contradiction matrix requires expert knowledge and manual effort, which is a limitation for novices such as students. Specifically, the main difficulties encountered by novices is problem abstraction, that is, the abstraction a specific engineering problem into a general problem, and solution specification, finding specific solutions from the abstract level (Hmina et al., 2024). It therefore limits TRIZ's accessibility and scalability in educational settings where the students' knowledge and experience of a new problem are very limited.

Recent advances in large language models (LLMs) have introduced new possibilities for supporting problem-solving and creative ideation across disciplines (Mohammadi & Zeng, 2025). These models excel at interpreting natural language and can therefore naturally support idea generation (e.g. Wang et al., 2023; Zhu & Luo, 2023). On the other hand, they lack the structured methodology needed for complex technical problem-solving.

While the TRIZ contradiction matrix requires users to find contradictions and apply its inventive principles manually, LLMs can help generate contradictions and solution suggestions autonomously or interactively in response to user prompts, see Figure 1. Together, the TRIZ contradiction matrix and LLMs they can therefore be complementary tools for modern design thinking (Čok et al., 2025).

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

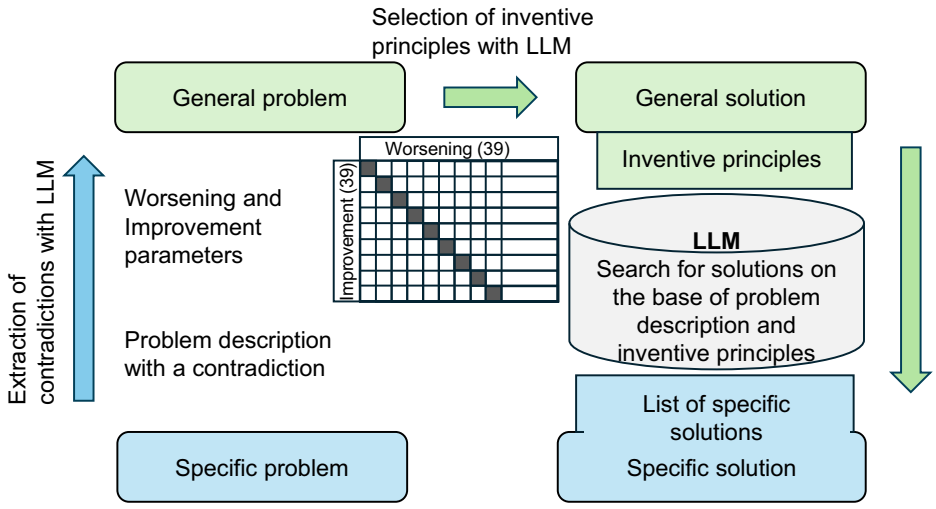


Figure 1. Proposed integration of LLM with the contradiction matrix method (TRIZ-C+LLM)

This study investigates how the use of the TRIZ contradiction matrix with the support of LLM-based artificial intelligence (AI) assistants, also called chatbots (hereafter shortened to TRIZ-C+LLM) can enhance design quality among engineering students. Students interacted with LLMs using prompts. Assessing the effectiveness of an enhanced method is usually not sufficient, other parameters are also important, depending on the method (Motte and Eriksson, 2016; Gericke et al., 2022). In this context, the extra workload potentially generated using LLM-based AI assistants, as well as the perceived usefulness and satisfaction for the TRIZ-C+LLM method were therefore also evaluated. Hence, this study aims at answering the following two research questions:

- RQ1: Does the use of TRIZ-C+LLM compared to the use of TRIZ contradictions alone (TRIZ-C) lead to improvements in design quality?
- RQ2: How do students perceive and evaluate the use of TRIZ-C+LLM in terms of workload, usefulness of the methods, and satisfaction in the results?

2. Literature review

TRIZ and LLMs are both deeply rooted in knowledge. TRIZ is based on a curated database of inventive principles, engineering contradictions, and patent analyses, while LLMs are trained on vast and diverse text corpora that include technical, scientific, and creative content. This shared reliance on structured knowledge enables both systems to support innovation and creativity. TRIZ uses abstraction and contradiction resolution to suggest inventive solutions, while LLMs rely on pattern recognition and probabilistic reasoning to generate novel ideas and analogies. Both approaches help expand the designer's thinking and unlock new possibilities (Mohammadi & Zeng, 2025).

Recent studies highlight the growing potential of LLMs to enhance TRIZ-based innovation and design. Chen, Song, et al. (2024) introduced a workflow that reformulates concrete problems into TRIZ-compatible formats, enabling structured reasoning and inventive solution generation. Jiang and Luo (2025) developed AutoTRIZ, an AI-driven tool that automates ideation within the TRIZ framework by leveraging LLMs' reasoning capabilities. Mohammadi and Zeng (2025) integrated TRIZ with Environment-Based Design (EBD) and LLMs to support early-stage design exploration emphasizing how LLMs can complement TRIZ by offering flexible, context-aware ideation. Arjomandi et al. (2024) emphasized the importance of tailored datasets for testing and validation, and Mas'udah and Livotov (2024) explored nature-inspired AI approaches to design. Recent work shows a clear convergence: AI is increasingly used to make TRIZ contradiction handling more precise, scalable, and context-aware. Trapp et al. (2023) demonstrate that transformer-based question-answering models can extract and formulate Inventive Design Methodology (IDM-TRIZ) contradictions directly from patent data, reducing ambiguity and reliance on expert interpretation. Complementing this, Mysior and Cavallucci (2025) show that LLMs can dynamically contextualize and reformulate TRIZ principles, overcoming the rigidity and sensitivity of the classical contradiction matrix. Together, these contributions illustrate how generative AI can operationalize contradiction reasoning more effectively in both formulation and

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

resolution stages. Despite these advances, several challenges remain. Interpretability of AI-generated TRIZ solutions is a key concern. Meincke et al. (2024) found that models like GPT-4 can outperform humans in idea generation speed and quality, but translating these ideas into actionable innovations still requires human expertise (Chiarello, 2024). AI-generated solutions often favour feasibility over novelty (Ma et al., 2023), and their effectiveness depends on contextual understanding and domain-specific integration.

Industry adoption also poses hurdles. Gmeiner et al. (2023) noted that designers often struggle to collaborate with AI tools, calling for new learning strategies. Xu et al. (2024) observed that while AI excels at knowledge extraction, its decision-making limitations necessitate human oversight. Gomez et al. (2024) and Ege et al. (2024) emphasized the need to balance AI-generated insights with human creativity in technical design. In previous studies engineering experts found both TRIZ and TRIZ+LLM useful, noting that LLM integration can accelerate ideation but may increase workload due to evaluating multiple generated options. Ease of use favours pure LLM approaches, but TRIZ+LLM offers structured depth (Jiang et al., 2025; Čok et al., 2025).

Use of LLMs to support ideation with novices are more promising. Engineering students find TRIZ complex and time-intensive to learn, with a heavy theoretical focus that limits motivation and practical application (Hmina et al., 2024). Meanwhile, students are already using LLM-based AI assistants frequently in their work and are satisfied with the results, even if they are not introduced into prompt engineering (Žáková, 2025). They are open to integration between LLMs and classical creativity and innovation methods. Chen, Tsang, et al. (2024) have shown promising results using an LLM-augmented morphological analysis approach. Results showed that the analogue group slightly outperformed the ChatGPT group in flexibility and originality. This suggests that a similar LLM-TRIZ integration could accelerate and support invention for novices. Still, the place and contribution of LLMs in ideation processes needs to be clearly established. Using LLMs for the sole purpose of increasing creativity should not be the only strategy. As a case in point, Bunn et al. (2025) compared student-team ideation with and without ChatGPT: the analogue group slightly outperformed the ChatGPT group in flexibility and originality. Where LLMs can also be useful is in reducing the mental effort required by TRIZ (problem abstraction and solution specification). Urban et al. (2024) showed that this was the case for idea generation. This in turn should lead to the students getting a better appreciation of the TRIZ method.

Finally, prior research shows that students tend to overestimate their solution when LLMs are used (Urban et al., 2024), which means that there can be an exaggerated trust in those systems. This potential caveat needs to be considered in such studies.

In summary, while LLMs and AI offer promising enhancements to TRIZ and engineering design, their integration is still emerging. Combining AI's computational power with TRIZ's structured methodology could accelerate innovation, but maintaining a balance between automation and human creativity, and taking into account the user's experience and appreciation of the process and the generate results, is essential.

3. Methodology

3.1. The engineering problems

The experiment was conducted on three different engineering problems that consisted in the development of solutions for:

- Polymer gear testing rig (GT): As part of a polymer gear testing rig, develop a concept for the alternating loading of polymer gears.
- Mounting bracket (MB): Develop a bracket for the mounting of a specified electrical drive onto a bicycle frame.
- Wind turbine power maximization (WT): Try to find concept for the maximization of wind turbine power (the power physical model equation of a wind turbine was given).

The problems, called GT, MB, and WT, differed in their degree of difficulty. In this context, difficulty is defined as the challenge of finding relevant solutions using general engineering knowledge. The GT and MB problems were relatively specialized and therefore required more expertise. The WT problem was more general and could be solved with general engineering knowledge.

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

3.2. Experimental setup and measures

3.2.1. Experimental setup

The experimental design was as follows. Three contrasting engineering problems were used, and all were solved either with TRIZ-C+LLM or with TRIZ-C only. The GT and MB problems were of similar difficulty and were relatively niched problems, requiring therefore more expertise. The WT problem was more general and required less expertise. The three problems were addressed individually by students. Each student solved one problem only and was randomly assigned the method to use, either TRIZ-C+LLM or TRIZ-C. The GT and MB problems were part of a course, and the students continued to work on the problem afterward. This made the experimental setup more realistic for the students and ensured motivation to solve those problems. The WT problem was not part of the course but due to of its more general nature, it was not expected to decrease motivation. The students working with GT were mainly in their fourth year of mechanical engineering studies. The students working with MB were Erasmus exchange master students and the majority had backgrounds in mechanical engineering. Finally, the students working with WT were in their fourth or fifth year of industrial engineering program but had a common mechanical engineering background (two years for the fourth-year students and three years for the fifth-year students). A specific characteristic of this last group was their specialization in AI, which made them more experienced in prompting (See Table 1).

The experiments were preceded by a pilot study in which the protocol was tested on the GT and WT problem using Microsoft Copilot with GPT-5 (see Figure 2), and showed correct contradictions and principles. As shown on Figure 2, Group 2 corresponds to the students solving the GT problem, Group 3 to the students solving the MB problems, Group 4 to the students solving the WT problem.

Table 1. Problems and students' technical background

Problem	Difficulty level	Student curriculum	Year of study	N1	N2
GT	Difficult and niched	ME (22)	4 (20), 5 (2)	21*	21*
MB	Difficult and niched	ME (12), IE (4), GE (3)	3 (3), 4 (6), 5(8), 6 (2)	20	19**
WT	Medium difficult and general	IE with AI (73)	4 (37), 5 (36)	73	73

(ME: mechanical engineering, IE: industrial engineering, GE: general engineering, N1: number of students performing the assignment, N2: number of students answering the questionnaire; number of students per year of study or curriculum in parenthesis).

*One student did not understand the engineering problem and was therefore excluded from the study.

**One student did not answer the survey.

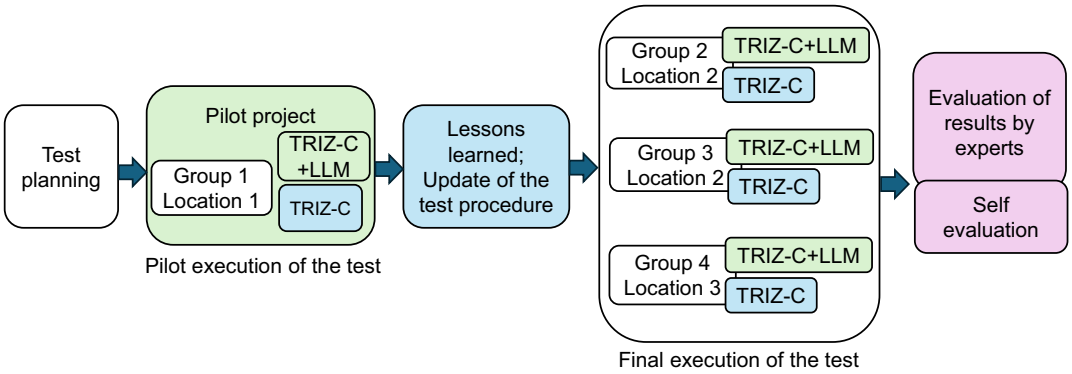


Figure 2. The flowchart of the conducted experiment

3.2.2. Design quality measure

Design quality was measured in the following four criteria:

- C1: Feasibility. How feasible is it to implement the solution?
- C2: Requirements. How well does the solution meet technical criteria and requirements?
- C3: Innovativeness of solution.
- C4: Elaboration. How elaborate (well-developed) is the solution?

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

The criteria were assessed on a seven-point scale (from 1 to 7) by the experts holding PhDs in mechanical engineering. Design quality was computed as the average of these four criteria. For the difficult problems MB and GT, two experts with prior experience in the problem domains, who were not personally involved in this design problem, independently evaluated the solutions, and the average of their scores was computed for each criterion. For the technically easier WT problem, another independent expert evaluated all the solutions.

3.2.3. Students' workload, usefulness and satisfaction measures

The student perception on workload, usefulness and satisfaction was obtained through an online post-experimental survey. The workload was assessed by using the relevant subscales of the NASA-Task Load Index (NASA-TLX, Hart, 2006) in original or slightly modified formulations. Questions or statements relative to the usefulness of the method, and satisfaction in the results can be found in Figure 4. No hypotheses were made about the outcomes of the survey questions or statements although the pilot study revealed that the workload was experienced as higher for the students using TRIZ-C+LLM.
The survey also contained control questions about the experiment (e.g., "I found the task instructions to be clear and easy to follow") and open questions about the used method. The students using TRIZ-C+LLM could leave their prompts and prompt answers as well.

3.3. Procedure of the experiment

All sites performed the same test protocol – with the exceptions marked below. Before the experimental tasks, the participants had a 30 min presentation of TRIZ and a 15 min of prompt engineering. They also completed a short practice problem to familiarize themselves with the procedure (TRIZ steps, and LLM use). The practice task was distinct from the test problems and not included in the analysis. The test had the following procedure planned to take about 45 minutes (the student could use more or less time) while the specific procedure for TRIZ-C and TRIZ-C+LLM respectively is presented Figure 3:

- Problem presentation.
- Search and selection of the relevant contradictions.
- In the WT problem, a contradiction and its related principles were subequently imposed to facilitate the design quality assessment. In the problems MB and GT the students could focus on the contradictions of their choice in order to not limit the solution space.
- The students then tried to find ideas based on at least three inventive principles for the WT problem or on their selected principles for the MB and GT problems,
- The students then further developed and documented their top-ranked idea.
- The participants answered individually to the survey.

The students using the TRIZ-C+LLM method were given a prompt-pack with concrete prompt instructions and examples for 1) problem framing (clarify the challenge), 2) TRIZ contradiction parameters identification, 3) relevant TRIZ inventive principles suggestions, 4) solution generation and 5) diversity check. They were recommended to use Microsoft Copilot with the LLM option Smart GPT-5 as the preliminary tests had shown that the GPT-5 correctly reported TRIZ contradictions and inventive principles. The students were however free to choose other LLM-based AI assistants, but had to report which one they had used. The students using TRIZ-C were given access to the https://www.triz40.com/ site and were asked to not use any LLM-based AI assistants.

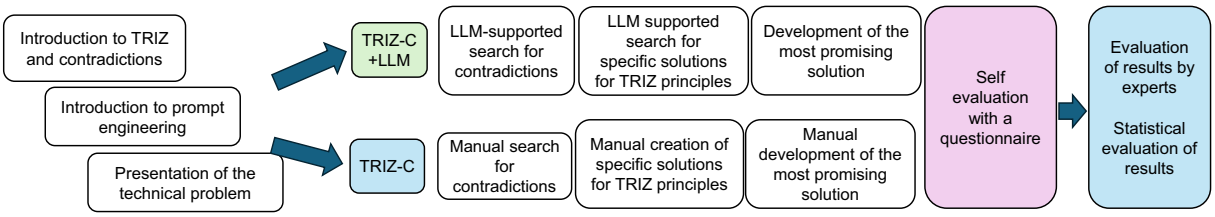


Figure 3. A detailed plan for the conducted test, comparison between TRIZ-C only and TRIZ-C+LLM

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

4. Results

All the reported results are on a seven-point scale (from 1 to 7) unless mentioned otherwise.

4.1. Experimental setup

The students evaluated the problems MB and GT as slightly more difficult than WT, with a mean of 4.2 and 4.1 vs. 3.5 ($F(2, 110)=2.59$, $p = .079$), as planned.
The average instruction clarity was 5. The instructions clarity regarding the TRIZ-C+LLM method for the WT problem was the only one to significantly deviate from the others with a mean of 5.7 (reason not identified). The time given to the students to perform the task was assessed as sufficient by the students with an average of 5, and not significant deviations across conditions.

4.2. Design quality

The obtained design quality means and standard deviations (SD) are given in Table 2. A two-way ANOVA was performed to investigate the main effects of the Problem and Method factors. The data points of the normal Q-Q plot of the standardized residuals of the design quality variable were close to the diagonal, indicating that data is normally distributed. The Levene's test rejected the null hypothesis of homogenous variance, $F(5, 108) = 2.88$, $p = .018$. Because of the uneven sample sizes and the small groups having larger variance, the F values were likely to be liberal (Tomarken & Serlin, 1986) but this did not affect the largely significant results ($p < .001$).
The main effect of the Problem factor was clearly significant, $F(2, 108) = 27.89$, $p < .001$, implying that there are differences in the design quality of the students' solutions depending on the problems. A post-hoc comparison using the Bonferroni test confirmed that the mean design quality for the WT problems significantly differed from the MB and GT problems ($p < .001$ for both comparison). The difference between MB and GT was also relatively large but not significant ($p = .095$). Regarding the main effect of the Method factor, the following values were obtained: $F(1, 108) = 3.21$, $p = .076$, which is not significant. Finally, there is a significant interaction between methods and problems, $F(2, 108) = 7.58$, $p = .001$, implying that methods different results for some problems. A further simple effect analysis revealed that the mean design quality score between methods (TRIZ-C or TRIZ-C+LLM) were not significant different for the GT problem, $F(1, 108) = .51$, $p = .479$, and for the MB problem, $F(1, 108) = .06$, $p = .809$. This was however the case for the WT problem, $F(1, 108) = 34.67$, $p < .001$.

Table 2. Mean and SD for the design quality and student satisfaction (in parenthesis) across problems and methods

Problem	Method	Mean	SD	N
GT	TRIZ-C	4.9 (4.4)	0.91 (1.75)	11
	TRIZ-C+LLM	4.6 (3.7)	1.17 (1.25)	10
	Total	4.7 (4.0)	1.03 (1.53)	21
MB	TRIZ-C	4.2 (3.7)	0.98 (1.25)	10
	TRIZ-C+LLM	4.2 (4.0)	0.93 (0.87)	10 (9)
	Total	4.2 (3.8)	0.93 (1.07)	20 (19)
WT	TRIZ-C	5.0 (4.5)	0.78 (1.36)	30
	TRIZ-C+LLM	6.2 (5.1)	0.62 (1.07)	43
	Total	5.7 (4.9)	0.88 (1.22)	73
Total	TRIZ-C	4.8 (4.3)	0.90 (1.44)	51
	TRIZ-C+LLM	5.6 (4.7)	1.13 (1.22)	63 (62)
	Total	5.3 (4.5)	1.10 (1.33)	114 (113)

4.3. Students' workload, usefulness and satisfaction

Beyond design quality, it was important to assess the students' workload, perceived usefulness and their own assessment of the design quality (satisfaction with the final result). Independent samples t-tests were

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

therefore conducted to compare the TRIZ-C+LLM group ($n = 62$) and TRIZ-C group ($n = 51$) across multiple outcome variables (Figure 4).

Regarding *workload*, the analysis revealed two statistically significant differences between the groups, both favouring the TRIZ-C+LLM condition. Participants in the TRIZ-C+LLM group reported significantly less difficulty in defining contradictions (M = 3.18, SD = 1.52) compared to the TRIZ-C group (M = 4.37, SD = 1.55), $t(111) = -4.123$, $p < .001$, $d = -0.779$. Similarly, the TRIZ-C+LLM group rated the overall problem as significantly less difficult (M = 3.31, SD = 1.36) than the TRIZ-C group (M = 4.27, SD = 1.34), $t(111) = -3.784$, $p < .001$, $d = -0.715$. The two effects demonstrated medium-to-large effect sizes. Regarding mental effort, only the MB and WT students answered and the results are inconclusive. There is an interaction between the problem and method factors, $F(1,88) = 6.89$, $p = .048$, with simple effects showing a significance difference between methods for the WT problem but not for the MB problem.

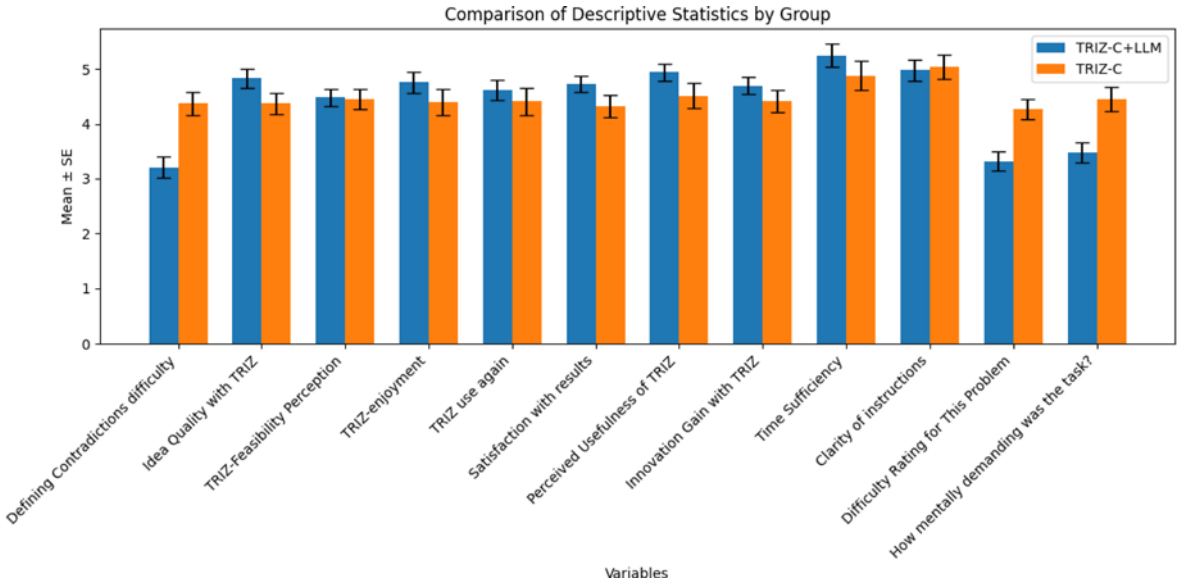


Figure 4. Comparison of descriptive statistics by group: TRIZ-C and TRIZ-C+LLM

Regarding *perceived usefulness*, no significant differences emerged for perceived usefulness ($p = .133$), willingness to use TRIZ again ($p = .541$), time sufficiency ($p = .307$), or clarity of instructions ($p = .766$). Task completion time was also comparable between groups (TRIZ-C+LMM: M = 40.3 minutes, SD = 20.9; TRIZ-C: M = 46.2 minutes, SD = 21.3). Noteworthy is that idea quality is strongly correlated with perceived usefulness (Spearman's $\rho = .684$, $p < .001$) and the "would-use-TRIZ-again" variable ($\rho = .687$, $p < .001$), implying that the obtained results are more influent in assessing usefulness than other characteristics the methods could have.

Regarding *satisfaction*, no significant differences emerged for innovation gain ($p = .321$), quality of the generated ideas ($p = .099$), feasibility perception ($p = .945$), enjoyment ($p = .271$), and satisfaction with results ($p = .119$). While there is no difference across methods, the experts and the students' evaluation of the satisfaction with the final concept correlated with $\rho(113) = 0.38$, $p < .001$. This correlation is high compared to similar studies (cf. Urban et al., 2024). Moreover, the average student satisfaction score were consistent with the experts' and systematically lower across all conditions (see Table 2). This means that the students have a generally accurate understanding of the quality of their results, although underestimating them.

5. Discussion

The results reveal significant differences in design quality depending on the difficulty level of the problems backgrounds. This parameter must therefore be carefully controlled in future experiments. The planned and perceived level of difficulty of the problem matched, which could make them suitable as benchmarks for similar studies. Contrary to initial expectations, the integration of LLMs did not consistently improve outcomes for difficult, specialized problems – precisely the cases where additional support was most needed. On the other hand, design quality was significantly higher for students using

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

TRIZ-C+LLM in the WT problem, which is a more general and less demanding task. Upon further analysis, two main reasons can be suggested. It is possible that LLMs may perform better when solutions can be derived from broadly available knowledge in their training corpus, while struggling with specialized contexts. Partially confirming this assumption, it has been reported by students that the LLMs sometimes misinterpreted the GT problem, focusing on gears rather than the gear testing rig. Another possible reason is that the students solving the WT problem, having a specialization in AI, were more experienced in prompting. Problem difficulty level, prompting experience, or both, can therefore explain this result. Further studies are necessary to disentangle this important aspect.

When designing the experiment, we anticipated that combining TRIZ with LLMs would yield a clear advantage over TRIZ alone. The findings show that when a problem is clearly defined and communicated to the LLM (e.g., through competent prompting), and the solution space can be easily explored by the LLM, useful results can be expected. This could explain the contradictory results from other studies in creativity research: analogue groups often outperform LLM-supported groups in flexibility and originality (Mohammadi & Zeng, 2025; Bunn et al., 2025; Ma et al., 2023), while other work demonstrates the potential of AI-driven tools to accelerate ideation (Jiang & Luo, 2025; Hmina et al., 2024; Urban et al., 2024).

Importantly, students' overall experience favoured TRIZ-C+LLM. The LLM-based AI assistant helped them identify contradictions and find specific solutions – activities often perceived as difficult by novices –, while the results on perceived usefulness and satisfaction were on a par with the TRIZ-C method.

The students' satisfaction with their results correlated with the experts' and the students' score were in average underestimated compared to the experts'. This is at odd with the current literature according to which novices tend to overestimate their skills (e.g., Kruger & Dunning, 1999). The fact that the students using LLMs did not overestimate their results, contrary to other results (Urban et al., 2024) is very positive as it means that the students' assessment of their solutions are not affected by the LLMs' sometimes very persuasive answers.

5.1. Limitations

This study was conducted with individual students rather than teams, which differs from the common practice in creative problem-solving tasks. While this approach limited the group dynamics that often foster idea generation, it provided clearer insights into individual competencies and how each student interacted with TRIZ and support of LLM AI assistant.

Another limitation concerns the complexity of the TRIZ methodology itself. Introducing TRIZ within a short timeframe – focused mainly on contradictions – posed challenges for students with limited prior exposure. Greater expertise in TRIZ would likely improve the reliability of the results. It also remains unclear whether the design solutions were primarily inspired by TRIZ principles or derived from general engineering knowledge. Nevertheless, both TRIZ-C and TRIZ-C+LLM groups had comparable levels of TRIZ knowledge, ensuring a fair comparison.

Differences in problem type, various LLM-based AI assistants, prompting skill, and student backgrounds likely influenced the results, but these factors were not controlled, introducing potential bias and making it unclear which variable had the greater impact. Using a single problem would have produced cleaner but highly case-specific findings, so three problems were used to provide a broader perspective. This approach offered a more holistic view but reduced experimental control.

Finally, prompting competence emerged as an important factor influencing outcomes in the TRIZ-C+LLM condition. Although students were provided with a prompt pack, differences in prior prompting experience may have affected performance. Since prompting skills were not treated as a control parameter, this introduces potential bias. Future studies should explicitly account for prompting competence, as well as problem difficulty, to better isolate the effects of LLM AI assistant support in design tasks.

6. Conclusions

This study examined the integration of LLM-based AI assistants into the TRIZ contradiction matrix (TRIZ-C+LLM) to support engineering students in creative problem-solving. Compared to the TRIZ contradiction matrix alone, LLM-based AI assistants did not consistently improve solution quality for complex or niche problems. However, LLM-based AI assistants helped students better understand contradictions, find specific solutions with clear benefits for more general tasks. The students'

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

evaluations correlated well with expert assessments, suggesting that students had a realistic, though slightly conservative, view of their solution quality. Taken together, these elements indicate that further studies on the TRIZ-C+LLM method are worth pursuing.

The findings highlight prompting experience as a critical factor. While students received a prompt pack, those with greater prior prompting experience may have had an advantage. Future experiments should control for this variable and investigate prompting practices in more detail. Further development of TRIZ-C+LLM should also investigate the accuracy of contradiction identification and inventive principle selection. Addressing these aspects could strengthen the method's reliability and broaden its applicability. Overall, the study underscores the importance of problem difficulty and prompting experience in shaping outcomes. Further research should control for these variables. The research points to TRIZ-C+LLM as a promising approach for enhancing both educational accessibility and practical innovation in engineering design.

Finally, the difficulty of the LLM-based AI assistants to help finding relevant and innovative solutions for difficult, niched problems shows that, currently, advanced tools or methods alone cannot guarantee high-quality results. Individual competencies – both in engineering knowledge and creative problem-solving – remain decisive.

Data availability

The data and materials used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgement

This research and its activities were supported by the Slovenian Research and Innovation Agency (ARIS) under the research program P2-0425: Decentralized solutions for the digitalization of industry and smart cities and communities; By the framework of the GREENTECH project, co-financed by the European Union – NextGenerationEU; Partly by the GenAID – Erasmus+ project contract number 2025-1-AT01-KA220-HED-000365381; And partly by the Lise Meitner professorship fund of the Faculty of Engineering LTH, Lund University, Sweden.

References

Arjomandi Rad, M., Hajali T., Martinsson Bonde J., Panarotto, M., Wärmefjord, K., Malmqvist, J., & Isaksson, O. (2024). Datasets in design research: needs and challenges and the role of AI and GPT in filling the gaps. *Proceedings of the Design Society*, 4, 1919–1928. https://doi.org/10.1017/pds.2024.194

Bunn, B.J., Howell, B.F., & Wright, G. (2025). Understanding ChatGPT's impact on student-team ideation outcomes for new product development. *Proceedings of the Design Society*, 5, 3281–3289. https://doi.org/10.1017/pds.2025.10342

Chen, L., Song, Y., Ding, S., Sun, L., Childs, P., & Zuo, H. (2024). TRIZ-GPT: an LLM-augmented method for problem-solving. In *ASME 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Article V006T06A010. ASME. https://doi.org/10.1115/detc2024-143163

Chen, L., Tsang, Y., Jing, Q., & Sun, L. (2024). An LLM-augmented morphological analysis approach for conceptual design. In C. Gray, E. Ciliotta Chehade, P. Hekkert, L. Forlano, P. Ciuccarelli, & P. Lloyd (eds.), *DRS2024: Boston*. Design Research Society. https://doi.org/10.21606/drs.2024.605

Chiarello, F., Barandoni, S., Majda Škec, M., & Fantoni, G. (2024). Generative large language models in engineering design: opportunities and challenges. *Proceedings of the Design Society*, 4, 1959–1968. https://doi.org/10.1017/pds.2024.198

Čok, V, Samsa, L, Brojan, M, Tavčar, J, & Vukašinović, N. (2025). Case study: is there a space for TRIZ in the era of ChatGPT? *Proceedings of the Design Society*, 5, 871–880. https://doi.org/10.1017/pds.2025.10101

Ege, D. N., Øvrebø, H. H., Stubberud, V., Berg, M. F., Steinert, M., & Vestad, H. (2024). Benchmarking AI design skills: insights from ChatGPT's participation in a prototyping hackathon. *Proceedings of the Design Society*, 4, 1999–2008. https://doi.org/10.1017/pds.2024.202

Gericke, K., Eckert, C., & Stacey, M. (2022). Elements of a design method – a basis for describing and evaluating design methods. *Design Science*, 8, Article e29. https://doi.org/10.1017/dsj.2022.23

Gmeiner, F., Yang, H., Yao, L., Holstein, K., & Martelaro, N. (2023). Exploring challenges and opportunities to support designers in learning to co-create with AI-based manufacturing design tools. In *CHI'23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 226. Association for Computing Machinery (ACM). https://doi.org/10.1145/3544548.3580999

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press

Gomez, A. P., Krus, P., Panarotto, M., & Isaksson, O. (2024). Large language models in complex system design. *Proceedings of the Design Society*, 4, 2197–2206. https://doi.org/10.1017/pds.2024.222
Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. https://doi.org/10.1177/154193120605000909
Hmina, K., Allouch, A., Bouyarmane, H., El Amine, M., & Sallaou, M. (2024). TRIZ effect exploitation on engineering students thinking skills in product design. *FME Transactions*, 52(2), 271–278. https://doi.org/10.5937/fme2402271H
Jiang, S., Li, W., Qian, Y., Zhang, Y., & Luo, J. (2025). AutoTRIZ: automating engineering innovation with TRIZ and large language models. *Advanced Engineering Informatics*, 65, *Part C*, Article 103312, https://doi.org/10.1016/j.aei.2025.103312
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121
Ma, K., Grandi, D., McComb, C., & Goucher-Lambert, K. (2023). Conceptual design generation using large language models. In *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Article V006T06A021. ASME. https://doi.org/10.1115/detc2023-116838
Mas'udah, M., & Livotov, P. (2024). Nature's lessons, AI's power: sustainable process design with generative AI. *Proceedings of the Design Society*, 4, 2129–2138. https://doi.org/10.1017/pds.2024.215
Meincke, L., Girotra, K., Nave, G., Terwiesch, C., & Ulrich, K. T. (2024). Using large language models for idea generation in innovation. *The Wharton School Research Paper Forthcoming*. https://doi.org/10.2139/ssrn.4526071
Motte, D., & Eriksson, M. (2016). Assessment framework for a methodology under development – Application to the PDA methodology. In *Proceedings of the 11th International Symposium on Tools and Methods of Competitive Engineering – TMCE'16* (pp. 373–388). Delft University of Technology. Retrieved November 12, 2025, from https://portal.research.lu.se/en/publications/assessment-framework-for-a-methodology-under-development-applicat/
Mohammadi, A., & Zeng, Y. (2025). Enhancing TRIZ through environment-based design methodology supported by a large language model, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 39, Article e12. https://doi.org/10.1017/S0890060425000083
Mysior, M., & Cavallucci, D. (2025). Contradiction Processing Using Large Language Models and Generative Artificial Intelligence. In *World Conference of AI-Powered Innovation and TRIZ Methodology. 2nd IFIP WG 5.4 International TRIZ Future Conference, TRAI 2025*. (pp. 171–183). Springer. https://doi.org/10.1007/978-3-032-08847-5_12
Orloff, M. A. (2020). *Modern TRIZ Modeling in Master Programs, Introduction to TRIZ Basics at University and Industry*. Springer Nature. https://doi.org/10.1007/978-3-030-37417-4
Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99(1), 90–99. https://doi.org/10.1037/0033-2909.99.1.90
Trapp, S., Großer, N., & Warschat, J. (2023). Question Answering with Transformers and Few-Shot Learning to Find Inventive Solutions for IDM-TRIZ Problems and Contradictions in Patents. In *Towards AI-Aided Invention and Innovation. 23rd International TRIZ Future Conference, TFC 2023* (pp. 23–42). Springer. https://doi.org/10.1007/978-3-031-42532-5_2
Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: an experimental study. *Computers & Education*, 215, Article 105031. https://doi.org/10.1016/j.compedu.2024.105031
Wang, B., Zuo, H., Cai, Z., Yin, Y., Childs, P., Sun, L. & Chen, L. (2023). A task-decomposed AI-aided approach for generative conceptual design. In *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Article V006T06A009. ASME. https://doi.org/10.1115/detc2023-109087
Zhu, Q., & Luo, J. (2023). Generative transformers for design concept generation. *Journal of Computing and Information Science in Engineering*, 23(4). https://doi.org/10.1115/1.4056220
Xu, W., Kotecha, M. C., & McAdams, D. A. (2024). How good is ChatGPT? An exploratory study on ChatGPT's performance in engineering design tasks and subjective decision-making. *Proceedings of the Design Society*, 4, 2307–2316. https://doi.org/10.1017/pds.2024.233
Žáková, K., Urbano, D., Cruz-Correia, R., Guzmán, J. L., & Matišák, J. (2025). Exploring student and teacher perspectives on ChatGPT's impact in higher education. *Education and Information Technologies*, 30(1), 649–692. https://doi.org/10.1007/s10639-024-13184-y

https://doi.org/10.1017/pds.2026.10409 Published online by Cambridge University Press