

## Article

# Dialogic Artificial Intelligence in the Architectural Design Studio: An Empirical Study of the AI Sparring Model

Mirko Stanimirovic <sup>1,\*</sup>, Jiaqi Wang <sup>2</sup>, Domen Zupancic <sup>3</sup>, Ana Momcilovic Petronijevic <sup>1</sup> and Ivan Ciric <sup>4</sup>

<sup>1</sup> Faculty of Civil Engineering and Architecture, University of Nis, 18000 Nis, Serbia; ana.momcilovic.petronijevic@gaf.ni.ac.rs

<sup>2</sup> College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China; wjqnancy0@zju.edu.cn

<sup>3</sup> Faculty of Architecture, University of Ljubljana, 1000 Ljubljana, Slovenia; domen.zupancic@fa.uni-lj.si

<sup>4</sup> Faculty of Mechanical Engineering, University of Nis, 18000 Nis, Serbia; ivan.ciric@masfak.ni.ac.rs

\* Correspondence: mirko.stanimirovic@gaf.ni.ac.rs

## Abstract

Integrating generative artificial intelligence into architectural education raises a question not only of whether AI helps, but of how its use should be structured. This paper empirically examines the AI Sparring model—a dialogic, critical approach to student–AI interaction—through a workshop in which architecture students designed a pavilion in a fortress. The three self-selected groups differed in the role AI played in the formation of the concept, not in its final presentation: Group A developed concepts traditionally, through sketches and physical working models, with AI serving only as a rendering tool at the end; Group B used AI directly as a generative tool; and Group C worked through structured critical dialogue with AI. The works were independently rated by three evaluators from three countries, alongside keyword analysis and assessment of reflective depth and theoretical references. Group C produced the most consistent concept quality, with no weak works, yet the single highest-rated work came from the traditional group—Sparring raises the “floor” rather than the “ceiling” of quality. At the same time, Group C reported the lowest sense of authorship, and qualitative analysis revealed “simulated depth”: theoretical references cited without genuine engagement. Because of the small, self-selected samples, the findings are read as exploratory and hypothesis-generating. The paper sheds light on the limits of reflective AI pedagogy and proposes concrete revisions to the model.

**Keywords:** artificial intelligence; architectural education; AI sparring; critical thinking; reflective practice; generative AI; design studio; authorship; AI colab

Academic Editors: Derek Clements-Croome and Antonio Caggiano

Received: 30 June 2026

Revised: 13 July 2026

Accepted: 17 July 2026

Published: 19 July 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

In earlier work, we reviewed the literature on architectural design, education, creativity, and artificial intelligence (AI) [1]. The conclusion was clear: there is a lack of practical testing of how much AI actually helps young architects—and of what different ways of using it produce. We therefore proposed the AI Sparring model, which, through provocation and reflection, is meant to drive the cognitive process in design. The idea builds on a shift from Schön’s [2] “learning by doing” to “learning by co-doing” where student and AI together shape reasoning through dialogue and re-examination.

Architectural design education is, moreover, only one strand of a much broader adoption of AI across the architecture, engineering and construction (AEC) sector. Ma-

chine learning and language models are increasingly used for construction inspection and pre-completion quality management—for example, for the automated classification and monitoring of fire-door defects [3,4]—as well as for AI-driven optimization of design schemes [5], generative modelling of architectural elements [6], and construction and facility management. At the same time, comparative studies report a persistent gap between academic research and industry practice in AI adoption [7,8]. Positioning AI Sparring within this wider landscape matters: the competencies the model targets—critical evaluation of AI output and its conscious integration into one’s own reasoning—are precisely those graduates will need in an industry where AI already permeates workflows far beyond design.

AI is increasingly present in architectural practice, and with it grows concern for the future of the profession and of education. Part of that concern is about automation and part of it is about cognitive deskilling: studies [9,10] warn that AI without a structured framework can worsen decision-making. It is precisely this question—not whether, but how to use AI—that motivates this paper.

To map current trends, we conducted a new literature search in ScienceDirect and Web of Science (2024–2026), using keywords such as “generative AI,” “architectural design,” “design pedagogy,” and “critical thinking.” Applying criteria of relevance to architectural design education, empirical or theoretical grounding, and peer review, we retained 33 papers; the full overview, with study types, samples, and relevance to the present study, is given in Appendix C (Table A5). The field is expanding rapidly: a recent bibliometric analysis maps a sharp acceleration of research on AI in conceptual design [11].

The selected works show several consistent trends. First, AI is most effective in the early phases of design—conceptualization, exploration, and ideation—while in later technical phases its role is, for some authors, more limited [12–14]. Second, students’ enthusiasm grows while teachers remain cautious; some authors stress the instructor’s key role in shaping a critical stance toward AI [15–17]. Third, the absence of a structured framework for interacting with AI leads to negative synergy in decision-making tasks [10,18]. Finally, a fourth group of authors argues that contemporary pedagogy increasingly recognizes the need for “learning by co-doing”—a process in which student and AI jointly shape design reasoning [19,20].

Iranmanesh [19] condenses these trends into four themes: phase-dependent integration of AI, new studio skills (prompting, evaluation, translating AI suggestions into decisions), pedagogical scaffolds for co-doing, and questions of authorship and ethics. All four concern AI Sparring: it targets the early phases, develops new skills, relies on a structured scaffold, and demands a conscious, critical relationship with AI. Phase-dependent integration points to where sparring may be most effective—in early design, when exploration, ideation, and critical re-examination are needed. The new studio skills are exactly what sparring develops: posing productive questions, assessing AI suggestions, and translating them into one’s own design decisions. Pedagogical scaffolds, such as structured dialogic sequences and time frames for reflection, form the basis of the workshop design. Questions of authorship and ethics, in turn, are central to a critical stance toward AI—sparring does not encourage blind acceptance but conscious re-examination of AI’s role in one’s own creative process.

It is worth noting that several of these works are empirically grounded. For example, Labib et al. [16] conducted a large survey ( $N = 239$ ) of students’ perceptions of generative AI in Saudi Arabia and Egypt. Ozorhon et al. [20] applied a design-based research methodology to integrating Midjourney into an architectural studio. Alshahrani & Mostafa [15] compared the attitudes of students and teachers in Saudi Arabia. Karasu & Yirmibeş [21] experimented with a hybrid approach combining physical models and generative AI in

real time. These works provide a solid empirical basis for understanding how students and teachers respond to AI in an educational context; similar findings are reported beyond architecture, in graphic-design education [17].

Although all of these works offer important insights, they rest mostly on surveys, short experiments, or conceptual frameworks. A recent systematic review of generative AI in architectural design education reaches the same conclusion, identifying the absence of empirical evidence on how AI use affects students' critical thinking as a key gap in the field [22]. What is missing are studies that follow structured, dialogic student–AI work across the whole process—exactly what Sparring proposes. The approaches closest to ours also structure the interaction, but they locate the structure elsewhere: in choreographed prompt workflows and curatorial decision-making [23], in instructor-led AI mentoring [24], or in framing AI as a creative agent under the studio's control [20]. AI Sparring places the structure in the dialogue itself, and makes it deliberately adversarial: the AI's role is to resist, provoke, and demand justification, so that cognitive friction rather than fluency drives the student's reasoning. The aim of this paper is to empirically examine, through a student workshop, the model's effects on the design process, creativity, and critical thinking. We start from the assumption that structured dialogue with AI can deepen critical thinking and metacognition more than uncontrolled AI use—and we compare the findings with the studies cited above.

## 2. Materials and Methods

We carried out the research in three phases: a workshop with students, analysis of their responses, and comparison with the literature. The aim was to test the AI Sparring model in practice and to identify patterns in how students work with AI.

### 2.1. The AI Sparring Model

AI Sparring is a dialogic way of working with generative AI, developed and argued in our earlier work [1]. Unlike ordinary use, where the AI's suggestion is taken as a finished result, Sparring treats AI as a critical interlocutor. The student poses a problem, the AI responds, and the student then analyses, critiques, and “returns the punch”. The goal is not a quick solution but deeper reflection through resistance.

Resistance is introduced through three “moves”: seeking error (the AI deliberately proposes a flawed solution and then defends it), contradiction (a solution that reconciles mutually exclusive demands), and an outside viewpoint (the AI argues from an opponent's perspective). For example, in the *seeking error* move, the AI proposes an all-glass pavilion on the fortress rampart and defends it in terms of “transparency and lightness”; the student's task is to identify why the proposal fails in this context—materiality, thermal behaviour, the dialogue with the stone walls—and to argue the critique with theoretical support. Further examples of the provocative prompts used are given in Appendix A (Table A2). After each round, the student writes a theory-based critical analysis, redefines the problem, and begins a new round.

The protocol used in this workshop had eight steps—from the initial intention, through two rounds of prompting and analysis, to the final concept and a reflective report (at least 500 words). Detailed instructions for Group C are in Appendix A, and the assessment guide is in Appendix B.

### 2.2. Research Design

The study is comparative, with three groups. Before the workshop, students chose their way of working based on the description on the website [25]. Group A worked traditionally—hand sketches and physical models; AI was permitted only for rendering the

final presentation. Group B used AI as a tool for quickly generating ideas. Group C applied AI Sparring—a structured, critical dialogue with AI.

The task was the same for everyone: to design a pavilion in the fortress. Students worked individually and submitted their works as PDFs. The focus was on the design process, not the final solution—we were interested in how the approach shapes thinking and decisions.

### 2.3. Participants

The call was published on the faculty website, with applications via Google Forms. We received 94 applications: 51 (54.3%) for Group A, 17 (18.1%) for B, and 26 (27.7%) for C. Already, here it is clear that the majority chose the traditional way of working, in which AI plays no role in concept formation, despite the workshop's theme (possible self-selection effects are discussed in the Discussion).

Not all applicants took part. We analysed 34 works—26 from A, 2 from B, and 6 from C. The survey was completed by 32 students (26 A, 2 B, and 4 C). This imbalance is a limitation of the study: results for B and C are read purely descriptively. Participants were undergraduate and master's students of various years.

The permitted AI tools were ChatGPT (OpenAI: GPT-5.5 Instant, built 23 April 2026.) and Gemini (Google: Nano Banana 2 and Gemini 3.1 Flash Image, built 26 February 2026.) used both for dialogue and for generating images through their built-in capabilities (students had free access for one month). No separate image-generation applications were used; all visual material was produced within these two tools. The traditional group used pencil, paper, and physical models—but in the end it too used AI for visualization: photographs of models were turned into renders. The defining boundary between the groups, therefore, concerns the process of concept formation, not the final artefact: in Group A, all design decisions were made through sketching and physical modelling, and AI entered only after the concept was complete, as a one-way visualization instrument—functionally equivalent to a rendering engine—whose output could not feed back into the design. The implications of this boundary are discussed further in the Discussion. The sample is summarized in Table 1.

**Table 1.** Overview of the sample by group.

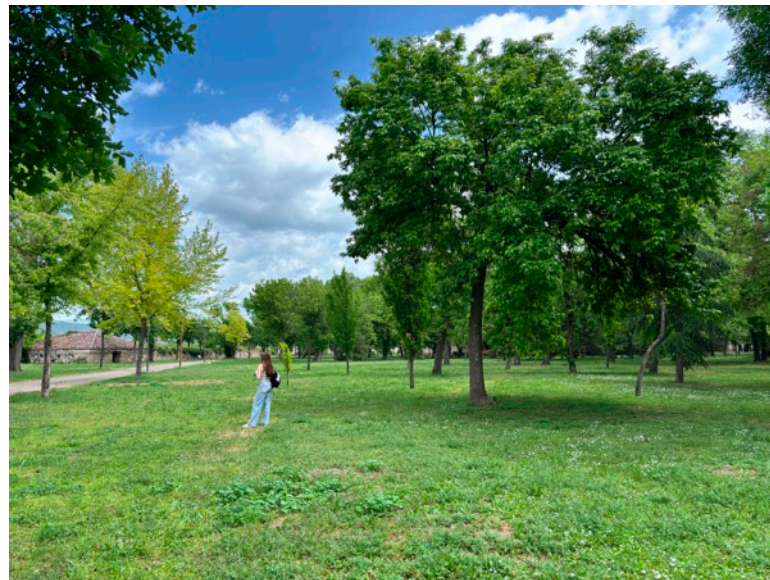
| Group            | Applications (Google Forms) | Submitted Works (Teams) | Survey Responses | Treatment Description                                               |
|------------------|-----------------------------|-------------------------|------------------|---------------------------------------------------------------------|
| A (traditional)  | 51                          | 26                      | 26               | Hand sketching and physical models; AI for final visualization only |
| B (AI as a tool) | 17                          | 2                       | 2                | AI generation + collage, no analysis                                |
| C (AI Sparring)  | 26                          | 6                       | 4                | Dialogue with AI, critical analysis, reflection                     |
| Total            | 94                          | 34                      | 32               | —                                                                   |

### 2.4. Procedure

The workshop ran from 18 to 27 May 2026. Three mentors (two from Serbia, one from Slovenia) held four three-hour meetings; between meetings students worked at home and submitted their solutions on Teams.

The first meeting was introductory, the second and third devoted to work and consultations, and the fourth to submission and the survey. The introductory meeting covered the task brief, the history and constraints of the site, the protocols of the three groups, and access to the permitted AI tools. During the two consultation meetings, mentors reviewed work in progress with each student; for Group C, this included reading the spar-

ring transcripts and probing whether the cited theoretical sources had actually been engaged with. Between meetings students worked at home, submitting all intermediate materials on Teams, which allowed the mentors to follow the process rather than only its outcome; the implications of working outside the studio environment are addressed in the Discussion. Twice the students visited the site in the fortress (Figure 1); photographs from the field were posted on the website for those who could not attend. Through the website, each participant received instructions, an anonymous ID (for unbiased grading), and Teams access. Instructions for Group C are in Appendix A.



**Figure 1.** Location for the pavilion in the fortress (photo: M. Stanimirovic).

### 2.5. Treatments by Group

Group A worked traditionally: three to five hand sketches, selection of the best, development through plan, section, and elevation, and a short concept description—using AI only for the final visualisation.

Group B used AI as a tool: a written initial intention, generating three to five variants by changing prompts and styles, selecting the most interesting, and a final collage, with a short description. The emphasis was on speed and visual exploration, without deeper analysis.

Group C applied AI Sparring following the protocol in Section 2.1: initial intention, a provocative prompt, three AI variants (one deliberately problematic), theory-based critical analysis, redefinition of the problem, a second round, and finally, an image of the pavilion generated from the refined concept, accompanied by a reflective report. The goal was not a quick solution but a deeper understanding of one's own thinking.

The three conditions form a deliberate spectrum of AI's role in conceptual work: in A, AI is absent from ideation and enters only at final visualisation; in B, AI is a direct generator used without critical mediation; in C, AI is a dialogic partner. Group B is thus essential to the design as the "uncontrolled AI use" condition against which Sparring is defined. In actuality, few students chose it ( $n = 2$ ); we retain it for this conceptual role but treat it illustratively throughout, drawing no statistical conclusions from it.

In all three groups, the process ended with an AI-generated image; what differed was its role—in Group A, a render of an already-completed traditional design; in Group B, the direct output of the request; and in Group C, the outcome of the preceding critical dialogue.

## 2.6. Evaluation

The works were independently rated by three evaluators—two mentors (Serbia and Slovenia) and an external expert (China)—using a predefined guide. The criteria were creativity and originality, concept consistency, feasibility, and presentation quality, and for Group C, also the quality of the dialogue and reflection. Grading was anonymous (by ID).

In assessing concept quality, evaluators were guided not by visual appeal alone but, following Pallasmaa [26], by a multisensory understanding of architectural experience that resists the visual primacy of much contemporary design. Each evaluator's score combined the predefined objective criteria (Appendix B) with their own informed perception, including emotional response and personal experience. This subjective component was deliberately bounded rather than eliminated: all three evaluators applied the same predefined criteria and anchored rubrics (Appendix B), rated independently and without mutual communication, and worked with anonymized submissions identified only by student ID. The level of agreement actually reached— $ICC(2,k) = 0.84$  for the averaged ratings—indicates that the shared criteria, rather than individual taste, dominated the scores.

*Note: All values for concept quality, reflective score, and theoretical references are the average of three evaluators' ratings (Serbia, Slovenia, and China). Inter-rater agreement for concept quality is good for the average measure ( $ICC(2,k) = 0.84$ ) and moderate for a single rater ( $ICC(2,1) = 0.64$ ).*

We collected both quantitative and qualitative data. Quantitative: concept quality (average of three evaluators), number of variants, number of reflective cycles (C), reflective depth (by rubric), and number of theoretical references. Qualitative: students' reflections, surveys, and sparring-dialogue transcripts (C).

Reflective depth was assessed with a five-point rubric adapted from Mezirow [27], from surface (1) to transformative reflection (5); the full rubric is in Appendix B. Theoretical references were counted manually: every mention of an author, concept, principle, period, style, or specific work was counted as a reference, and works were placed on a 0–5 scale (Appendix B).

## 2.7. Data Collection

Data come from several sources: student works (sketches, collages, prompts, and reports) on Teams as PDFs, and surveys in Google Forms. Everything was consolidated into a Google Sheets database with columns for ID, group, survey average, concept quality, number of iterations, reflective score, and theoretical references (Table 2). For Group C, dialogue transcripts were additionally collected, which enabled the keyword analysis.

**Table 2.** Structure of the data collected per student.

| Column                 | Description                                           |
|------------------------|-------------------------------------------------------|
| Student ID             | Anonymous code (e.g., A-01, B-01, C-01)               |
| Group                  | A (traditional), B (AI as a tool), C (AI Sparring)    |
| Survey average         | Mean score on a Likert scale (1–5)                    |
| Concept quality        | Mean rating of independent evaluators (1–5)           |
| Number of iterations   | Number of design cycles/AI interactions (for B and C) |
| Reflective score       | Rubric-based reflective depth (1–5)—Group C only      |
| Theoretical references | Number of theoretical references in the text (0–5)    |

During the preparation of the original text, the authors used AI tools (DeepSeek and Claude) solely for language polishing and translation of the original Serbian text into English. The content was then reviewed and edited, and the authors take full responsibility for the final manuscript.

### 3. Results

We analysed concept quality, reflective depth, and the number of theoretical references, together with a qualitative reading of the reports and surveys. Given the size of Groups B ( $n = 2$ ) and C ( $n = 6$ ), no inferential tests were performed, and group-level summary statistics (mean, SD) are reported only for Group A ( $n = 26$ ). For Groups B and C, all quantitative results are presented at the level of the individual student, and comparisons between groups are made by direct counts rather than by summary statistics.

#### 3.1. Quality of the Conceptual Solution

Concept quality was rated by three independent evaluators (Serbia, Slovenia, and China) on a 1–5 scale; inter-rater reliability was high ( $\text{ICC}(2,k) = 0.84$ ). Group A's scores are summarised in Table 3; the individual scores of all eight students in Groups B and C are given in Table 4.

**Table 3.** Concept-quality scores for Group A (1–5;  $n = 26$ ).

| Group           | N  | Mean | SD   | Min  | Max  |
|-----------------|----|------|------|------|------|
| A (traditional) | 26 | 3.63 | 0.83 | 2.00 | 5.00 |

*Note: SD is the standard deviation (the spread of scores around the mean—smaller means more uniform quality within the group). Values are the average of the three evaluators' ratings. Dispersion statistics are reported only for Group A; Groups B and C are too small for meaningful dispersion statistics and are reported individually in Table 4.*

Every work in Group C was rated 4.00 or higher (individual averages 4.00–4.67), while the two Group B works received 3.67 (B-55) and 4.67 (B-70). The ordering of group levels—Group C consistently above Group A—is confirmed by all three evaluators, including the external evaluator from China. Group B's two works are shown not as a statistical comparison but because Group B embodies the direct, uncritical AI use against which Sparring is defined, and they should be read in that light throughout.

**Table 4.** Individual results for all students in Groups B and C. Each value is the average of the three evaluators' ratings; the number of iterations is a recorded count.

| Student ID | Concept Quality (1–5) | No. of Iterations | Reflective Score (1–5) | Theoretical References (0–5) |
|------------|-----------------------|-------------------|------------------------|------------------------------|
| B-55       | 3.67                  | 3                 | —                      | —                            |
| B-70       | 4.67                  | 3                 | —                      | —                            |
| C-3        | 4.67                  | 3                 | 4.33                   | 4.67                         |
| C-17       | 4.33                  | 3                 | 1.67                   | 0.00                         |
| C-18       | 4.33                  | 3                 | 2.33                   | 3.00                         |
| C-45       | 4.33                  | 4                 | 3.33                   | 0.00                         |
| C-49       | 4.67                  | 3                 | 4.67                   | 4.33                         |
| C-65       | 4.00                  | 5                 | 2.33                   | 0.00                         |

*Note: Reflective reports were required only in Group C; '—' means the measure was not collected for that group.*

Group C's position does not come from a higher “ceiling”. The single best work in the study is from Group A (rated 5.00 by all three evaluators), and A has five works rated  $\geq 4.5$ —as many as the best works in C. The difference lies at the bottom of the distribution (Table 5): in Group A, 15 of 26 works fall below 4.0 and five below 3.0, whereas no work in Group C falls below 4.0. Sparring, then, is not associated with a higher peak but with the absence of weak works.

**Table 5.** Distribution of concept-quality scores by group (number of works per band).

| Group            | <3.0 | 3.0–3.99 | 4.0–4.49 | ≥4.5 |
|------------------|------|----------|----------|------|
| A (traditional)  | 5    | 10       | 6        | 5    |
| B (AI as a tool) | 0    | 1        | 0        | 1    |
| C (AI Sparring)  | 0    | 0        | 4        | 2    |

### 3.2. Depth of Reflection

Reflective depth was assessed with a 1–5 rubric, and only for Group C, since they alone wrote reflective reports. The six individual scores (Table 4) cluster in two bands: three students at or below 2.33 (C-17, C-18, C-65) and three at or above 3.33 (C-3, C-45, C-49), with no student in between—a polarization that is directly visible in the individual data.

### 3.3. Theoretical References

The number of theoretical references (0–5 scale; concept descriptions and reports) is likewise reported per student in Table 4. For Groups A and B there were not enough data for a separate analysis (they wrote no reflective reports); their descriptions are discussed in Section 3.5. In Group C, the pattern is strictly bimodal: three students used no theoretical references at all (scored 0 by all three evaluators: C-17, C-45, C-65), while the remaining three (C-3, C-18, C-49) average between 3.00 and 4.67—with no middle ground. The shape of the pattern holds for every rater: the evaluator from China scored each of the three referencing students one to three points lower than the evaluators from Serbia and Slovenia, so her stricter scoring lowers the level of the referencing band but does not change the bimodal structure.

### 3.4. Number of Iterations and Survey Analysis

The number of design cycles differed by group. Group A averaged 1.23 iterations (SD 0.51; range 1–3). Both Group B students completed exactly three iterations. In Group C, four students completed three iterations, one completed four (C-45), and one completed five (C-65)—every Sparring student thus went through at least three full cycles of prompting, analysis, and reflection.

The survey was completed by 32 students (26 from A, 2 from B, 4 from C). All questionnaires shared five common questions (authorship, demandingness, willingness to repeat, learning about one’s own thinking, and satisfaction); the rest were group-specific. The common questions are shown in Table 6; given two respondents in B and four in C, the values for these groups are descriptive and should be read with corresponding caution.

**Table 6.** Average scores on the common survey questions (1–5).

| Question (Common to All Groups)           | A (N = 26) | B (N = 2) | C (N = 4) |
|-------------------------------------------|------------|-----------|-----------|
| I feel like the true author of my concept | 4.38       | 3.50      | 2.75      |
| The process was demanding but useful      | 4.38       | 3.50      | 2.75      |
| I would gladly repeat this way of working | 4.00       | 3.50      | 4.50      |
| I learned something about my own thinking | 4.12       | 4.00      | 4.75      |
| Overall satisfaction with the workshop    | 4.15       | 4.50      | 4.00      |

Group C rated the process lowest on “demanding but useful” (2.75) and highest on learning about its own thinking (4.75)—sparring was experienced as taxing but stimulating. The strongest sense of authorship is in Group A (4.38) and the weakest in Group C (2.75), with B in between (3.50). It is telling that C—whose works were rated most consist-

ently high by three evaluators—reports the weakest sense of authorship; we discuss this tension in the Discussion. Because of the small samples and differences in the questions, these findings are illustrative.

### 3.5. Qualitative Insights

Analysis of the reports (C) and concept descriptions (A and B) revealed several patterns. Group A described the process as “intuitive” and “personal,” but rarely with theory. The reflections are mostly descriptive—about what was drawn, not why. Descriptions come down to form and aesthetics (rhythm, dynamics, solid and void, and context), without authors or precedents. This corresponds to a score of one or two.

Group B focused on the aesthetics of the AI images. The descriptions are short and superficial, without reflection on the relationship between the AI’s suggestions and the students’ own decisions.

Group C showed the deepest reflection and more often used theory. One student linked an AI-proposed glass facade, incompatible with the stone fortress, to Frampton’s critique and “critical regionalism,” and then emphasized stone in the second round. Another described how sparring forced them to reconsider how uncritically they accept AI. Such statements indicate critical and transformative reflection (score 4–5).

### 3.6. Keyword Analysis

Keyword analysis (concept descriptions and initial intentions) shows (Table 7) that the most frequent categories, across all three groups, are nature/context (92 occurrences) and form/geometry (88 occurrences). The numbers are the total term occurrences per category in each group. These two categories often go together: in 23 of 33 students ( $\approx 70\%$ ), both appear. By group, a clear difference emerges. Group A deals with form (form/geometry 71, versus 46 for context). Group C is turned toward context (nature/context 36; form only seven). In Group B, the two are roughly equal (10 each).

**Table 7.** Keyword frequency by category and group. Values are absolute numbers of keyword occurrences in the students’ initial-intention and concept-description texts; group totals reflect group size (A: 26, B: 2, and C: 6).

| Category           | Group A | Group B | Group C |
|--------------------|---------|---------|---------|
| Nature/context     | 46      | 10      | 36      |
| Form/geometry      | 71      | 10      | 7       |
| Movement/dynamics  | 32      | 5       | 5       |
| Solid/void/open    | 30      | 4       | 7       |
| Material/structure | 21      | 4       | 4       |
| Light/shadow       | 21      | 2       | 1       |
| Rhythm/repetition  | 22      | 0       | 0       |

*Note:* Values are absolute numbers of keyword occurrences in the students’ initial-intention and concept-description texts; group totals reflect group size so comparisons should be read within groups or as proportions.

The vocabulary of rhythm and repetition appears only in Group A (22 occurrences). The theory/author/precedent category is close to zero in all three groups (A: one work; B: one; C: none in the initial intentions). An important distinction from the theoretical-reference score in Section 3.3 is that score also covers Group C’s later critical phase, the only phase where theory appears for some students—not in the initial intention. All parameters are summarized in Table 8.

**Table 8.** Comparative overview of results by group.

| Parameter                           | Group A ( <i>n</i> = 26) | Group B ( <i>n</i> = 2) | Group C ( <i>n</i> = 6)                        |
|-------------------------------------|--------------------------|-------------------------|------------------------------------------------|
| Concept quality (1–5)               | 3.63 (SD 0.83)           | 3.67; 4.67              | 4.00–4.67 (all $\geq$ 4.0)                     |
| Number of iterations                | 1.23 (SD 0.51)           | 3; 3                    | 3–5                                            |
| Reflective score (1–5)              | n/a                      | n/a                     | 1.67–4.67 (two bands)                          |
| Theoretical references (0–5)        | n/a                      | n/a                     | 0 (three students); 3.00–4.67 (three students) |
| Learned about own thinking (survey) | 4.12                     | 4.00                    | 4.75                                           |

*Note:* Group A: mean (SD); Groups B and C: individual values or range (see Table 4). Concept quality, reflective score, and theoretical references are the average of three evaluators' ratings (Serbia, Slovenia, and China); ICC(2,k) = 0.84 for concept quality. 'n/a' means the datum was not collected for that group. Survey values for Groups B and C rest on two and four respondents, respectively (see Table 6).

Group C has the highest values for concept quality, number of iterations, and self-assessed learning, with the most theoretical references and the deepest reflection (measured only for this group). At the same time, its advantage in quality lies in consistency, not in the peak (see Section 3.1). Group A gives moderate results, the strongest sense of authorship, and the widest range—from the weakest to the single best work. Group B is too small and not uniform (one work leans toward C's register). Overall, the data support the view that Sparring is associated with more uniformly high-quality concepts, deeper reflection, and richer theoretical grounding—but the conclusion remains descriptive because of the small, self-selected samples. We discuss the meaning of these findings in the Discussion. To illustrate the qualitative character of each group's process, Figures 2, 3, and 4 present a representative example from Group A, Group B, and Group C, respectively.

**Figure 2.** Student work A-47 (Group A).



**Figure 3.** Student work B-70 (Group B).



**Figure 4.** Student work C-65 (Group C).

## 4. Discussion

The results provide an empirical basis for understanding how structured dialogue with AI (Sparring) affects design, creativity, and critical thinking. Because the groups are small, unequal, and self-selected, we read the findings descriptively and as hypothesis-generating—as patterns to be tested, not as proof. Where the data challenge the model’s assumptions, we say so openly, because precisely those points say the most about what remains to be done.

### 4.1. AI Sparring and “Learning by Co-Doing”: Theoretical Framework

Our starting framework is close to “learning by co-doing” [19], which carries Schön’s [2,28] “learning by doing” into the age of AI. With AI, “doing” ceases to be purely human and becomes co-doing—the joint shaping of reasoning through dialogue. Vaccaro [10] also gives a reason for structure: without a framework, human–AI collaboration can yield negative synergy. The architectural design process is, moreover, not merely an efficiency task in spatial problem-solving: it demands contextualization—the ability to reason toward disruptive solutions that remain aligned with the essence of the task and the author’s own position.

Group C's process broadly matches co-doing: students asked, critiqued, redefined, and wrote reflections. But the picture is not unified—for part of the group, the outward signs of theoretical engagement did not match real depth (see Section 4.5). We therefore use the concept of co-doing as an interpretive frame, not as a claim that it was achieved for everyone.

Co-doing can be understood as an intermediate stage in the use of AI, where it functions as a tool with general response capabilities. In contrast, AI sparring represents a more advanced approach, engaging with research processes aimed at the development of AI agents. Within this framework, iterative design and decision-making processes may serve as reverse-engineering mechanisms, supported by integrated methods.

#### *4.2. Floor Rather than Ceiling: Consistency of Concept Quality*

Every Group C work was rated at least 4.0, while more than half of Group A's works fall below that level (Tables 3–5). The direction of the difference—Group C above Group A—is given by all three evaluators, including the external one from China ( $ICC(2,k) = 0.84$ ), which lends the pattern stability. This convergence is more telling than it first appears. Because scoring deliberately combined the defined criteria with each evaluator's own multisensory and emotional reading of the work (Section 2.6; [26]), the agreement of three assessors from three countries points to a shared, substantive perception of quality rather than an artefact of rubric-following.

Yet C's advantage is not a higher peak. The best work in the study is from Group A (rated 5.00 by all three evaluators); the maximum in Group C is 4.67. The difference is at the bottom: in A, 15 of 26 works are below 4.0 and five below 3.0, while in C none is below 4.0 (Table 5). Sparring, then, raises the “floor”, not the “ceiling”.

This fits the role of a pedagogical scaffold: the structure (provocation, contradiction, and a second round) helps those who would otherwise develop a shallow concept, while the best reach the goal even without it. Vendrell & Johnston [29] call this “cognitive friction”, and Karadağ & Ozar [30] link more complex prompting to a better concept. The finding complements Alamasi et al. [12]: what matters is not only that AI is used in the conceptual phase, but how. The closest comparative evidence points the same way. In a workshop of the same size ( $N = 36$ ), Uhrík et al. found that generative AI supports ideation most effectively when embedded in structured workflows that emphasize curatorial decision-making [23]. Thus, the same conclusion is reached using different structures: theirs organizes the workflow and the curation of outputs, ours the dialogue itself. Two caveats remain: part of the uniformity may be due to the small sample and the coarse scale, and the groups were self-selected (see Section 4.7). Group B ( $n = 2$ ) is reported illustratively; its two works differ widely (3.67 and 4.67), so these data do not allow the conclusion that structured AI outperforms ordinary use.

#### *4.3. Framing the Problem Through Language: Keyword Analysis*

Keyword analysis is, in our view, the methodologically strongest part of the study, because it is qualitative and independent of sample size. It shows a clear difference in how the groups frame the same task. Group A students mostly framed the problem through form—the form/geometry category is the most frequent there (71 occurrences), and the vocabulary of rhythm and repetition appears exclusively among them (22 occurrences). Group C, by contrast, frames the problem through its relationship to context and nature—the nature/context category strongly dominates (36 occurrences), while form/geometry is weakly represented.

This linguistic difference is consistent with the idea that structured interaction with AI directs the student's attention toward relational and contextual aspects of the problem, rather than only the object's formal properties—which can be linked to applications of

creativity theory to prompting [31] and to the student's role as the one who assesses and integrates AI suggestions [32]. It is important, however, to stress what this analysis measures and what it does not: it shows a difference in vocabulary and framing, but does not by itself prove greater theoretical depth. A further reason for caution is specific to how Group C worked. Because their concepts were ultimately realised as AI-generated images, theoretical and contextual vocabulary entered prompts that the tools answered with stylistic approximations rather than with reasoned design positions. The lexical turn toward context and nature may therefore reflect, in part, an adaptation to what the tool rewards, not only a deeper theoretical consciousness—a possibility consistent with the simulated depth discussed in 4.5. As the following sections show, the presence of contextual or theoretical vocabulary need not mean corresponding cognitive work behind it.

#### *4.4. The Authorship Paradox*

The most telling pattern is the tension between externally rated quality and the sense of authorship. Group C, with the highest and most consistent ratings from three evaluators, has the lowest sense of authorship: 2.75, versus 4.38 in A and 3.50 in B. The students who (with AI) produced the best works feel the least that the works are theirs. A second, more material explanation may complement this. Group C students observed that the AI images were not too crude but too finished: they arrived pre-loaded with materiality, lighting, and incidental detail no one had requested. As one student put it, “the lighting and tectonic detail were not what I had in mind—I only wanted to test the massing”. In the conceptual phase, the openness of a sketch or rough model is precisely what lets a student claim it as their own; an over-determined image fills that space in advance and resists internalisation. On this reading, the weak sense of authorship is in part a response to the visual surplus of the medium, not only to the cognitive cost of collaboration.

We interpret this contrast with a caveat—the survey for C was completed by four students. But the pattern is not dissatisfaction: the same group rates highest on learning about its own thinking (4.75) and on willingness to repeat the work (4.50). They recognize the value of the experience but do not experience the result as fully their own.

The finding connects to the theme of authorship and ethics [19] and introduces a tension into the “Designer Arbiter” role [32]: Group C students did select and assess AI suggestions, but without a stronger sense of ownership. Labib et al. [16] show that ethical awareness is an important factor; our data suggest that intensive co-creation with AI can weaken the sense of authorship even when quality is high. Part of the explanation, as in 4.5, may be that references and decisions were passed on from the AI rather than internalized—so the low sense of authorship is not only a cost of collaboration but also an accurate impression that part of the work is not theirs.

#### *4.5. The Limits of Reflection: Simulated Depth*

Analysis of the works and conversations with Group C students revealed a phenomenon that calls part of the model's assumptions into question. We call it simulated depth: all the outward signs of theoretical engagement are present—names, precedents, concepts—without the cognitive work behind them.

An example is the student who cited Bernard Tschumi as a precedent. Asked what she had noticed in his works, she said ChatGPT had suggested him and admitted she had not even looked at his projects. The reference was ornamental, not load-bearing. The case would not matter were it not part of a pattern: across all consultations, no Group C student raised with a mentor the topic of the specific examples or authors the AI had suggested. The same is independently noted by the second mentor (from Slovenia), who likewise did not see students studying the proposed examples more deeply. The agreement of two mentors gives the finding a weight a single case would not have.

This is reinforced by how the tools produce such references. Under hallucination and a limited context window, a name the model offers—“Tschumi” here—can be a statistically probable continuation rather than a grounded inference about that architect’s thought. The reference the student inherits may thus have no reasoning behind it on either side: the model did not derive it, and the student did not examine it. This gives a technical basis for treating AI-proposed theory with scepticism, and it compounds the methodological point that follows.

There are two consequences. The first is methodological: counting theoretical references—otherwise a strong instrument for us—measures the presence of vocabulary, not understanding, so the high values for part of Group C (3.3) should be read more cautiously. The second is a consequence for the model: Sparring is meant to prevent uncritical acceptance of AI, yet here uncritical acceptance disguised as critique was recorded—and in the group most trained to critique. The ritual of sparring—provocation, contradiction, and a second round—can be performed without the cognitive work. The “cognitive friction” [29] is here bypassed.

This pattern has a close analogue outside architecture. In a randomized experiment on learning with ChatGPT, Fan et al. describe “metacognitive laziness”: learners offload the regulatory work of learning onto the tool, improving immediate task performance without corresponding gains in knowledge [33]. Simulated depth can be read as the design-studio variant of the same mechanism—the sparring protocol supplied the outward form of critical engagement, while the regulatory work behind it was, for part of the group, delegated to the tool. Consistently with this, Xu et al. find that generative-AI environments support self-regulated learning only when metacognitive support is explicit rather than assumed [34], and comparisons of AI and human tutors suggest that current chatbots reproduce the form of Socratic questioning more readily than its epistemic pressure [35]. These findings anticipate the design conclusion below: the structure of the dialogue is not sufficient; the protocol must demand evidence of engagement.

Since one author is the creator of the model and a mentor, these observations are participatory observation—a valuable insight, but also a possible source of bias, which we state openly. Simulated depth was not the rule: some reports seem genuine, like the one in which a student links the misfit of a glass facade to a theoretical critique of context:

*“The AI proposed a glass facade that has nothing to do with the stone fortress. It reminded me of Frampton and his critique of ‘critical regionalism’—architecture should dialogue with context, not ignore it.”*

So, it is not that there was no depth—it was uneven and at times simulated, and our instrument cannot tell the two apart. This is confirmed by Group C’s bimodal reflective score: three students with a mean  $\leq 2.33$  and three with  $\geq 3.33$  (3.2). Part of the explanation may lie in the cognitive posture that image generation invites. When a fully rendered proposal appears in seconds, the student’s task shifts from divergent exploration—building a concept from nothing against an open field of possibilities—to the evaluation of an outcome that already exists. The slow unfolding of thought that characterises sketching or physical modelling is compressed, and reflection is displaced into a post hoc reaction to the AI’s output rather than emerging within the design process itself. A critique produced from this evaluative posture can carry all of the outward marks of engagement while bypassing the generative work that gives it substance.

This is not only a limitation but also a design lesson. If reflection can be performed through a statement, the protocol must require evidence of engagement, not a claim of it: an attachment of what the student actually looked at (a screenshot of the work, with a couple of sentences on what surprised them) or a brief oral defence of each reference. That way “I looked at Tschumi” becomes a verifiable step. The point where the model failed

becomes a contribution—a concrete way in which reflective AI pedagogy misses its target is described, together with a proposed fix.

#### 4.6. Implications for Practice and Model Design

With these caveats, cautious implications follow. Sparring offers a structured framework that, at least for some students, links AI with reflection rather than with a shortcut. The consistency pattern (4.2) suggests that the model is most valuable precisely for students at risk of a shallow concept—which future research can test directly. And simulated depth shows that structure is not enough: the protocol must build in a check of real engagement.

Indirectly, the “traditional” —“digital” boundary is not sharp: Group A too used AI in the end to visualize models. Sparring could therefore be offered as one option in the studio, not as a replacement for existing ways of working. A further design lesson concerns regional and historical authenticity. Asked to evoke local or traditional character—directly relevant on a fortress site—generative tools tend to reduce a rich tradition to a handful of signifiers and to conflate distinct periods and places into a generic “traditional” pastiche. A student without first-hand knowledge of the local tradition can mistake this stereotype for an authentic reference and build on it uncritically. Simulated depth, then, may obscure not only theoretical engagement but historical and regional authenticity, and a future protocol should require students to verify the sources and representativeness of any “traditional” element the AI proposes rather than accept it as given.

*Note: This concern was raised by the external evaluator, drawing on the Chinese context, who noted she could not directly assess the local authenticity of the Serbian works.*

#### 4.7. Limitations and Directions for Future Research

The study has serious limitations, which we state openly. The greatest is self-selection. Students chose their group themselves, so more motivated students may have gone for Sparring. This design does not separate the effect of the model from the effect of selection—no finding may be read causally. Moreover, the majority (54.3%) chose the traditional way of working, which further shows that group composition is not random. The preference itself deserves interpretation: in a workshop explicitly devoted to AI, more than half of the applicants (54.3%) chose to work without it. Several non-exclusive explanations are plausible. Before an evaluated task, students may rationally prefer the skills they have already mastered. The traditional format also promises undivided ownership of the result—consistent with Group A’s highest sense of authorship (4.38) and, conversely, with the erosion of that sense under intensive co-creation (Section 4.4). Ambivalence documented in the literature may contribute as well: surveys report divided views among architecture students about the risk of imitation and loss of originality [36], and AI awareness itself can act as a stressor as much as a motivator [37]. Whatever the mix, the pattern is a finding in its own right: voluntary adoption of AI-intensive formats cannot be assumed even among the generation presumed most receptive, and any curricular integration of models like Sparring must reckon with this baseline preference. Relatedly, where Group C’s critical thinking did appear to progress, this should be read with caution: with a small, self-selected sample we cannot rule out that some of it reflects chance or guessing rather than a stable gain. Second, the samples are small and unequal (A  $n = 26$ , B  $n = 2$ , C  $n = 6$ ; the C survey was completed by four students), so the quantitative findings are descriptive and comparisons involving B are illustrative. Even so, part of the contribution—the keyword analysis and simulated depth—does not depend on sample size.

Third, the measure of theoretical depth registers vocabulary, not understanding (4.5). Fourth, the dual role of one author (creator of the model and mentor) is a source of insight

but also of possible bias; it is mitigated by independent ratings from three evaluators in three countries ( $ICC(2,k) = 0.84$ ) and the agreement of two mentors on simulated depth.

Finally, the study was conducted at a single faculty, in a short workshop, and largely outside the traditional studio environment: between the four joint meetings, students worked at home, submitting intermediate materials on Teams. This reduced the spontaneous peer exchange and immediate mentor feedback that a studio normally provides, and although the Teams submissions allowed us to follow the process, we cannot exclude that the setting affected the quality and character of the works—in either direction. Generalization is therefore limited. We therefore understand the paper as exploratory and hypothesis-generating: it does not prove the superiority of Sparring but presents patterns—raising the floor, the authorship paradox, and simulated depth—worth testing. The next step would be a replication with random assignment to groups, a larger sample, and a protocol that requires evidence of engagement, not merely a statement. That such a design is feasible in this field has meanwhile been demonstrated: a study published during the revision of this paper used a cluster-randomized, three-group comparison across two cohorts ( $N = 145$ ) to separate the contribution of generative AI from that of general digitalization in architectural design teaching [38].

Several concrete devices could give that protocol teeth. Future iterations might move from general-purpose chatbots to instructor-configured agents whose workflow forces a counter-argument or contextual questions before any proposal, embedding opposition in the tool rather than leaving it to the student to construct. Submissions could include a decision log recording which choices changed under AI opposition and whether the student internalised the change—shifting assessment from the quality of the final image to the integrity of the decision path, and guarding against Sparring ending, as the traditional group's process did, in a final act of AI visualisation. Evaluation might also weigh not the number of iterations but their effectiveness, since many rounds can add cognitive load without deepening the concept. Such designs presuppose more than a short workshop: they require dedicated instructional time and basic AI literacy if students are not to treat the agent as a faster search engine.

Models that expose their chain-of-thought (e.g., DeepSeek) offer a further avenue—and a further risk. Their visible reasoning could serve as fresh Sparring material for the student to interrogate; but a fluent, seemingly rigorous chain may also invite the student to mistake the AI's logical completeness for design completeness and to transcribe its reasoning as their own. Reflective reports should therefore ask students to separate the AI's reasoning from their own explicitly—a safeguard against simulated depth migrating from cited references to borrowed logic.

## 5. Conclusions

This exploratory study tested the AI Sparring model in a real architectural design workshop and yielded a mixed, partly unexpected picture. The hypothesis—that structured dialogue with AI deepens critical thinking and yields more authentic work than uncontrolled AI use—is neither confirmed in a strong sense nor refuted; with small, self-selected groups, this design cannot do either. What the data do support is encouraging: Group C produced the most consistent concept quality, with no weak works—a pattern that held for each of the three independent evaluators.

But three results ran against expectation, and they are the most interesting. First, Sparring's advantage lies not in a higher ceiling—the single best work came from the traditional group—but in raising the floor. Second, the group with the best externally rated work reported the lowest sense of authorship, an authorship paradox rather than the expected reinforcement. Third, the depth the model aims for was often only simulated: the ritual of critique can be performed without the cognitive work behind it. These are not

failures to be hidden but the study's most original contributions, and they sharpen the model rather than discard it.

Architectural design will increasingly be carried out with AI rather than alongside it, which makes it urgent to ask not whether the technology is used but how it is integrated into the formation of design thinking. In traditional studio work—our Group A—it is the mentor who guides students toward reflection, supplying the theory, architectural history, and lateral, “out-of-the-box” thinking that an AI tool does not volunteer. Where AI is used without that depth, as in Group B, there is a risk that it displaces rather than supports critical reasoning: the results may be competent yet converge toward a shared, generic register, as the tool returns what is statistically typical rather than what is particular to a site or an author. This is not a danger confined to the studio; the same dynamic will shape practice, as clients and architects alike reach for AI to realise their intentions. Against this background, the Sparring model—our Group C—suggests one possible response: not the rejection of AI, but its deliberate use as a provocation, so that the friction of disagreement keeps the student's own design reasoning active rather than outsourced.

Yet this remains a proposal, not a proof. We present the work as hypothesis-generating, not as evidence of superiority. The findings justify a larger, randomized replication with a protocol that requires verifiable evidence of engagement with AI-proposed references. Read this way, the model's weak points become a concrete research agenda for reflective AI pedagogy.

**Author Contributions:** Conceptualization and methodology, M.S.; investigation, M.S., D.Z. and A.M.P.; formal analysis, M.S. and J.W.; validation, M.S., J.W. and D.Z.; writing—original draft preparation, M.S.; supervision, A.M.P. and I.C.; writing—review and editing, M.S., J.W. and D.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, under the Agreement on Financing the Scientific Research Work of Teaching Staff at the Faculty of Civil Engineering and Architecture, University of Nis—Registration number: 451-03-34/2026-03/200095 dated 5 February 2026.

**Institutional Review Board Statement:** Ethical review and approval were waived for this study, as it analyses anonymized student works and questionnaire responses from a voluntary, extracurricular educational workshop. Participants were informed of the research purpose through a public notice on the workshop website prior to enrolment, and all data were anonymized through ID codes.

**Informed Consent Statement:** Informed consent was obtained from all participants. A public notice on the workshop website, displayed alongside the enrolment form, stated that participation was voluntary and that works and questionnaire responses would be analysed and published for research purposes; voluntary enrolment under these conditions constituted consent.

**Data Availability Statement:** The data supporting the findings (anonymized student works, reflective reports, and summary tables) are available on reasonable request from the corresponding author, subject to restrictions protecting participant privacy.

**Acknowledgments:** During the preparation of this manuscript, the authors used DeepSeek and Claude for the purposes of language polishing and translation of the original Serbian manuscript into English. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** One author is the creator of the AI Sparring model and a mentor in the workshop, which is stated in the text as a possible source of bias; to mitigate this, independent rating was carried out by three evaluators from three countries. The authors declare no other conflicts of interest.

## Appendix A. Instructions for Group C (AI Sparring)

The following instructions were given to Group C students before they began. They operationalize the AI Sparring model from Section 2.1. The student poses a problem, the AI responds—but not always as expected—and the student analyses, critiques, and “returns the punch.” The procedure has eight steps (Table A1).

**Table A1.** Steps of the AI Sparring procedure.

| Step | What the Student Does                                                              | Time   |
|------|------------------------------------------------------------------------------------|--------|
| 1    | Writes the initial intention before first addressing the AI.                       | 10 min |
| 2    | Sends a provocative prompt (techniques in Table A2).                               | 5 min  |
| 3    | The AI generates three variants, one of them deliberately problematic.             | —      |
| 4    | Critical analysis: what the AI got right and what it got wrong, drawing on theory. | 20 min |
| 5    | Redefines the problem based on the analysis.                                       | 10 min |
| 6    | Second round: a new prompt and repeated analysis.                                  | 20 min |
| 7    | Creates the final concept (collage or sketch).                                     | 30 min |
| 8    | Reflective report (at least 500 words).                                            | 30 min |

Instead of neutral requests (e.g., “generate a modern pavilion”), students were encouraged to use provocative prompts that deliberately introduce resistance (Table A2).

**Table A2.** Examples of provocative prompts.

| Technique               | Example Prompt                                                                                                                  |
|-------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Seek error              | “Generate a pavilion that deliberately ignores the historical context, then explain why someone would defend such an approach.” |
| Introduce contradiction | “Design a pavilion that is at once fully open and fully closed, and justify the contradiction.”                                 |
| Take an outside view    | “Imagine you are an architect who dislikes this site—what pavilion would you propose, and why?”                                 |

In the end, the student submits a single PDF: the initial intention, both rounds of prompts with AI responses and analysis, the final concept, and the reflective report.

## Appendix B. Assessment Guide

The works were independently rated by three evaluators, blinded to group membership. Concept quality was rated on a 1–5 scale (1 = poor, 5 = excellent). Reflective depth was assessed with a rubric adapted from Mezirow [27] —Table A3.

**Table A3.** Rubric for assessing reflective depth (1–5).

| Score | Label          | Description                                                                                                                                                               |
|-------|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1     | Surface        | The student describes what they did, without analysing why; no dilemmas or change of mind.                                                                                |
| 2     | Descriptive    | States what was good or bad, without elaboration; mentions one dilemma but does not develop it.                                                                           |
| 3     | Analytic       | Identifies a discrepancy between intention and outcome and links it to one theoretical principle.                                                                         |
| 4     | Critical       | Analyses how sparring changed their thinking; two to three references (metacognition).                                                                                    |
| 5     | Transformative | Describes how the exercise changed their understanding of architecture or their own role; connects it to authorship, ethics, and the architect’s role in the digital age. |

Theoretical references were counted manually and classified by the scale in Table A4. A reference means a mention of an author, concept, principle, period, style, or specific work.

**Table A4.** Scale of theoretical references (0–5).

| Score | Criterion                                                                                                                                                          |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0     | No theoretical references; description only.                                                                                                                       |
| 1     | One superficially mentioned reference, without elaboration.                                                                                                        |
| 2     | Two to three references connected to one’s own concept.                                                                                                            |
| 3–5   | Three or more references, critically connected to decisions; within this band, the score reflects the number of distinct references identified (5 = five or more). |

Evaluators were instructed to assess the depth of thinking, not the style or length: a short text can be deep (4–5) and a long one superficial (1–2). The value in the summary table (Table 4) is the average of three evaluators’ ratings.

## Appendix C. Overview of the 33 Selected Studies

**Table A5.** Studies identified in the literature search (ScienceDirect and Web of Science, 2024–2026; keywords: “generative AI”, “architectural design”, “design pedagogy”, and “critical thinking”) and retained after applying the criteria of relevance, empirical or theoretical grounding, and peer review.

| No. | Main Author<br>(First Surname) | Year | Journal                                                 | Relevance to the Present Study                                                                                                                                |
|-----|--------------------------------|------|---------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | Abrusci et al. [39]            | 2025 | <i>Computers and Education: Artificial Intelligence</i> | Moderate. Even a short-structured AI intervention yields measurable creativity gains (2–12%, $p = 0.0003$ ; SUS = 77).                                        |
| 2   | Alamasi et al. [12]            | 2026 | <i>Ain Shams Engineering Journal</i>                    | High. Test group with GenAI outperformed the control, especially in form and aesthetics (14%); evidence that AI aids the conceptual phase.                    |
| 3   | Alana et al. [14]              | 2026 | <i>Buildings</i>                                        | High. Stage-based analysis (Spearman, Cronbach $\alpha = 0.78$ , Friedman $p < 0.05$ )—AI strongest in the conceptual stage and presentation.                 |
| 4   | Alshahrani & Mostafa [15]      | 2025 | <i>Frontiers in Built Environment</i>                   | High. Student enthusiasm vs. teacher scepticism; usage frequency and skill development as strongest predictors.                                               |
| 5   | Alymani et al. [8]             | 2026 | <i>Architecture</i>                                     | Contextual. AI/ML adoption in academia vs. industry; organizational capacity as the key determinant—background for institutional readiness.                   |
| 6   | Asfour [22]                    | 2026 | <i>Digital</i>                                          | High. Identifies the lack of empirical evidence on GenAI’s effects on students’ critical thinking as the field’s key gap—the gap the present study addresses. |
| 7   | Bloch et al. [7]               | 2025 | <i>Architectural Science Review</i>                     | Contextual (AEC). Research–practice gap in ML adoption; positions studio experiments as a bridge between academia and industry.                               |
| 8   | Chen et al. [11]               | 2026 | <i>Architecture</i>                                     | Moderate. Maps the acceleration of AI research in the conceptual design stage; situates the study within field-level trends.                                  |
| 9   | Çiçek et al. [40]              | 2025 | <i>Journal of Computational Design</i>                  | Moderate. AI as evaluator (VLM + ResNet-18, Grad-CAM); complements—not replaces—human judgement in assessment.                                                |
| 10  | Dharmatanna et al. [36]        | 2025 | <i>Modul</i>                                            | Moderate. AI used mostly for data gathering, concept narration and rendering; divided views                                                                   |

|    |                                        |      |                                                                       |                                                                                                                                                                                 |
|----|----------------------------------------|------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    |                                        |      |                                                                       | on imitation risk—a risk Sparring addresses through conscious evaluation.                                                                                                       |
| 11 | Ding et al. [37]                       | 2025 | <i>Acta Psychologica</i>                                              | Contextual. “Dual effect” of AI awareness (exhaustion vs. intrinsic motivation), moderated by AI literacy; relevant to how structured use re-frames AI-related stress.          |
| 12 | Hikmet & Ozay [41]                     | 2026 | <i>Buildings</i>                                                      | High. Schön (reflection-in/on-action) and algorithmic thinking as a framework for creative cognition in collaboration with AI.                                                  |
| 13 | Iranmanesh [19]                        | 2026 | <i>International Journal of Technology and Design Education</i>       | High. “Learning by co-doing” and four meta-themes (phase-dependent integration, new studio skills, pedagogical scaffolds, authorship/ethics)—Sparring operationalizes co-doing. |
| 14 | Jin et al. [42]                        | 2024 | <i>Buildings</i>                                                      | Moderate. AI-embedded teaching model; TAM-based evaluation framework (PU, PEU, ATU) for AI in education.                                                                        |
| 15 | Karadağ & Ozar [30]                    | 2025 | <i>Frontiers of Architectural Research</i>                            | High. Prompt strategies (keywords, connectives, progressive) linked to concept quality; baseline and longitudinal measurement.                                                  |
| 16 | Karasu & Yirmibeş [21]                 | 2026 | <i>Ain Shams Engineering Journal</i>                                  | High. Physical model + GenAI in real time; reflection-in-action through simultaneous feedback (476 interventions).                                                              |
| 17 | Labib et al. [16]                      | 2025 | <i>Sustainability</i>                                                 | High. Ethical awareness and perceived benefits predict skills and intention to use—empirical basis for the ethics dimension of Sparring.                                        |
| 18 | Lee & Kang [31]                        | 2025 | <i>Journal of Asian Architecture and Building Engineering</i>         | High. Boden’s creativity theory (combinatorial, exploratory, transformational) applied to prompting—structured interaction fosters creativity.                                  |
| 19 | Liu et al. [43]                        | 2025 | <i>Journal of Open Innovation: Technology, Market, and Complexity</i> | Moderate. Seven prompt components and eight techniques; transferable to the analysis of Sparring prompt “moves”.                                                                |
| 20 | Lobato-Valdespino & Flores-Romero [44] | 2024 | <i>Jornadas sobre Innovación Docente en Arquitectura</i>              | Moderate. “Extended intelligence” (Clark & Chalmers) as theoretical grounding for extended cognition in text–image design dialogue.                                             |
| 21 | Medel-Vera et al. [45]                 | 2025 | <i>Computers and Education: Artificial Intelligence</i>               | High. GenAI as collaborator and provocateur; instruments for measuring creativity (semantic analysis, aesthetic score, CSI) in AI-supported learning.                           |
| 22 | Obeidat [46]                           | 2026 | <i>Social Sciences and Humanities Open</i>                            | Moderate. Student perceptions (TAM), verification, ethics and cultural representation in a non-studio course.                                                                   |
| 23 | Ozorhon et al. [20]                    | 2025 | <i>International Journal of Technology and Design Education</i>       | High. AI as a “creative agent” integrated into the studio while retaining critical control; methodological reference for studio-embedded research.                              |
| 24 | Poturak [47]                           | 2025 | <i>Limes plus</i>                                                     | Contextual. AI not yet integrated into CAD/BIM workflows in a practice-changing way; education as the natural site for experimentation.                                         |
| 25 | Samaty & Albadi [24]                   | 2026 | <i>Alexandria Engineering Journal</i>                                 | High. Three-phase AI Mentoring Method; empirical support for the decisive role of the instructor in structured AI integration.                                                  |

|    |                           |      |                                                         |                                                                                                                                                                                                     |
|----|---------------------------|------|---------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 26 | Sopher & Varinlioglu [48] | 2025 | <i>eCAADe 2025 Proceedings</i>                          | Moderate. Process-level analysis of how AI supports problematization (future scenario planning, back-thinking).                                                                                     |
| 27 | Uhrík et al. [23]         | 2026 | <i>Architecture</i>                                     | High. Structured human–AI dialogue (“prompt choreographies”) with qualitative 0–5 evaluation; closest comparative study—structure located in workflow and curation rather than cognitive resistance |
| 28 | Vaccaro et al. [10]       | 2024 | <i>Nature Human Behaviour</i>                           | High. Human–AI combinations without structure can yield performance losses (negative synergy)—the core rationale for a structured protocol.                                                         |
| 29 | Vendrell & Johnston [29]  | 2026 | <i>Computers and Education: Artificial Intelligence</i> | High. Six essential critical-thinking processes and eight principles for preserving “cognitive friction”—direct theoretical grounding of the Sparring model.                                        |
| 30 | Wang & Yang [5]           | 2025 | <i>Procedia Computer Science</i>                        | Contextual (AEC). AI-driven design-scheme optimization (93.31% accuracy); technical AEC application outside pedagogy.                                                                               |
| 31 | Wang et al. [13]          | 2025 | <i>Buildings</i>                                        | High. Dual-module framework (AI skills + ethics); phase-dependent use—students most active in pre-design and design development.                                                                    |
| 32 | Yang & Wang [6]           | 2025 | <i>Machine Learning with Applications</i>               | Contextual (AEC). Generative modelling of traditional decoration; illustrates the technical AEC stream without a pedagogical dimension.                                                             |
| 33 | Yu [32]                   | 2025 | <i>Design Studies</i>                                   | High. “Designer Arbiter” role—the student as evaluator and integrator of AI suggestions; cognitive load in ideation and evaluation.                                                                 |

Note: Studies are listed alphabetically by the surname of the first author.

## References

1. Stanimirovic, M.; Momcilovic Petronijevic, A.; Stoiljkovic, B.; Kondic, S.; Nikolic, B. AI Sparring in Conceptual Architectural Design: A Systematic Review of Generative AI as a Pedagogical Partner (2015–2025). *Buildings* **2026**, *16*, 488. <https://doi.org/10.3390/buildings16030488>.
2. Schön, D. *Educating the Reflective Practitioner*; Jossey-Bass: Hoboken, NJ, USA, 1987.
3. Wang, S. Graph Neural Network–Driven Text Classification for Fire-Door Defect Inspection in Pre-Completion Construction. *Sci. Rep.* **2025**, *15*, 44382. <https://doi.org/10.1038/s41598-025-28120-4>.
4. Wang, S. Development of an Automated Transformer-Based Text Analysis Framework for Monitoring Fire Door Defects in Buildings. *Sci. Rep.* **2025**, *15*, 43910. <https://doi.org/10.1038/s41598-025-27648-9>.
5. Wang, X.; Yang, T. Optimization of Architectural Design Scheme Based on Artificial Intelligence Algorithm. *Procedia Comput. Sci.* **2025**, *261*, 990–997. <https://doi.org/10.1016/j.procs.2025.04.491>.
6. Yang, J.; Wang, G. Structure-Aware Stable Diffusion for Traditional Architectural Decoration Design. *Mach. Learn. Appl.* **2025**, *22*, 100735. <https://doi.org/10.1016/j.mlwa.2025.100735>.
7. Bloch, T.; Hamo-Goren, C.; Grobman, Y.J.; Austern, G. Unraveling the Impact of Machine Learning on the AEC Sector: Academia and Industry Perspectives. *Archit. Eng. Des. Manag.* **2026**, in press. <https://doi.org/10.1080/17452007.2026.2632103>.
8. Alymani, A.; Alsofiani, M.; Mandou, S.; Alubaidan, Z.; Al Tuwaijri, N. Artificial Intelligence and Machine Learning Implementation Patterns in Architecture: A Cross-Sectional Analysis of Academic and Industry Sectors in Saudi Arabia. *Architecture* **2026**, *6*, 57. <https://doi.org/10.3390/architecture6020057>.
9. Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* **2025**, *15*, 6. <https://doi.org/10.3390/soc15010006>.
10. Vaccaro, M.; Almaatouq, A.; Malone, T. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>.

11. Chen, L.; Chen, Z.; Dong, F. Adopting Artificial Intelligence in Architectural Conceptual Design: A Systematic Bibliometric Analysis. *Architecture* **2026**, *6*, 60. <https://doi.org/10.3390/architecture6020060>.
12. Alamasi, R.; Asfour, O.S.; Ashmeel, R. The Impact of Generative AI on Architectural Design Education: Insights from Hands-on Experience with Architecture Students. *Ain Shams Eng. J.* **2026**, *17*, 103879. <https://doi.org/10.1016/j.asej.2025.103879>.
13. Wang, J.; Shi, Y.; Chen, X.; Lan, Y.; Liu, S. Teaching with Artificial Intelligence in Architecture: Embedding Technical Skills and Ethical Reflection in a Core Design Studio. *Buildings* **2025**, *15*, 3069. <https://doi.org/10.3390/buildings15173069>.
14. Alana, H.; Fikry, M.; Hasan, A. Human–AI Collaborative Design in Architectural Studios: Evaluating Paradigm Shifts Across the Six Stages of the Design Process. *Buildings* **2026**, *16*, 1445. <https://doi.org/10.3390/buildings16071445>.
15. Alshahrani, A.; Mostafa, A.M. Enhancing the Use of Artificial Intelligence in Architectural Education—Case Study Saudi Arabia. *Front. Built Environ.* **2025**, *11*, 1610709. <https://doi.org/10.3389/fbuil.2025.1610709>.
16. Labib, W.; Abdelsattar, A.; Abowardah, E.; Abdelalim, M.; Mahmoud, H. The Role of Generative AI in Architecture Education from Students' Perspectives—A Cross-Sectional Descriptive and Correlational Study. *Sustainability* **2025**, *17*, 10029. <https://doi.org/10.3390/su172210029>.
17. Alsswey, A. Examining Students' Perspectives on the Use of Artificial Intelligence Tools in Higher Education: A Case Study on AI Tools of Graphic Design. *Acta Psychol.* **2025**, *258*, 105190, doi:10.1016/j.actpsy.2025.105190.
18. Drapal, A. Role of Memes in the Communication Process. *RISUS-J. Innov. Sustain.* **2025**, *16*, 60–76. <https://doi.org/10.23925/2179-3565.2025v16i2p60-76>.
19. Iranmanesh, A. Reflection in Co-Action: AI Joins the Crit. *Int. J. Technol. Des. Educ.* **2026**, in press. <https://doi.org/10.1007/s10798-026-10079-6>.
20. Özorhon, G.; Nitelik Gelirli, D.; Lekesiz, G.; Müezzinoğlu, C. AI-Assisted Architectural Design Studio (AI-a-ADS): How Artificial Intelligence Join the Architectural Design Studio? *Int. J. Technol. Des. Educ.* **2025**, *35*, 1999–2023. <https://doi.org/10.1007/s10798-025-09975-0>.
21. Karasu, B.M.; Yirmibeş, A. Real-Time Interaction of Physical Models and Generative AI in the Architectural Design Process (Hybrid Design Studio Experiments–HDS). *Ain Shams Eng. J.* **2026**, *17*, 104198. <https://doi.org/10.1016/j.asej.2026.104198>.
22. Alamasi, R.; Asfour, O.S. Applications of Generative AI in Architectural Design Education: A Systematic Review and Future Insights. *Digital* **2026**, *6*, 6. <https://doi.org/10.3390/digital6010006>.
23. Uhrík, M.; Cervantes, J.C.L.; Morales, C.E.S.; Hajtmanek, R.; Demčák, J.; Kupko, A. Prompt Choreographies: Dialogues Between Humans and Generative AI in Architecture. *Architecture* **2026**, *6*, 46. <https://doi.org/10.3390/architecture6010046>.
24. El Samaty, H.S.; Albadi, N. From Policy Frameworks to AI Mentoring Practice: A Structured Approach to Responsible Innovation in Architectural Education. *Alex. Eng. J.* **2026**, *137*, 206–217. <https://doi.org/10.1016/j.aej.2026.01.012>.
25. Mirko Stanimirovic AI Sparring. Available online: <https://wireframe22.wixsite.com/aisparring> (accessed on 30 June 2026).
26. Pallasmaa, J. *The Eyes of the Skin: Architecture and the Senses*; Wiley: Hoboken, NJ, USA, 2005.
27. Mezirow, J. *Transformative Dimensions of Adult Learning*; Jossey-Bass: San Francisco, CA, USA, 1991.
28. Schön, D. *The Reflective Practitioner: How Professionals Think in Action*; Basic Books: New York, NY, USA, 1984.
29. Vendrell, M.; Johnston, S.-K. Scaffolding Critical Thinking with Generative AI: Design Principles for Integrating Large Language Models in Higher Education. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100572. <https://doi.org/10.1016/j.caeai.2026.100572>.
30. Karadağ, D.; Ozar, B. A New Frontier in Design Studio: AI and Human Collaboration in Conceptual Design. *Front. Archit. Res.* **2025**, *14*, 1536–1550. <https://doi.org/10.1016/j.foar.2025.01.010>.
31. Lee, S.; Kang, S. The Influence of Generative AI with Prompt Engineering on Creative Design in Architectural Education. *J. Od Asian Archit. Build. Eng.* **2025**, *1*–16. <https://doi.org/10.1080/13467581.2025.2552446>.
32. Yu, W.F. AI as a Co-Creator and a Design Material: Transforming the Design Process. *Des. Stud.* **2025**, *97*, 1–27. <https://doi.org/10.1016/j.destud.2025.101303>.
33. Fan, Y.; Tang, L.; Le, H.; Shen, K.; Tan, S.; Zhao, Y.; Shen, Y.; Li, X.; Gašević, D. Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance. *Br. J. Educ. Technol.* **2025**, *56*, 489–530. <https://doi.org/10.1111/bjet.13544>.
34. Xu, X.; Qiao, L.; Cheng, N.; Liu, H.; Zhao, W. Enhancing Self-regulated Learning and Learning Experience in Generative AI Environments: The Critical Role of Metacognitive Support. *Br. J. Educ. Technol.* **2025**, *56*, 1842–1863. <https://doi.org/10.1111/bjet.13599>.
35. Fakour, H.; Imani, M. Socratic Wisdom in the Age of AI: A Comparative Study of ChatGPT and Human Tutors in Enhancing Critical Thinking Skills. *Front. Educ.* **2025**, *10*, 1–12. <https://doi.org/10.3389/feduc.2025.1528603>.

36. Dharmatanna, S.W.; Wijaya, E.S.; Pramananda, I.P.B. Integrating Ai in Architectural Design Thinking: An Opportunity and Challenges Study in Indonesia Architectural Education. *MODUL* **2025**, *25*, 55–63. <https://doi.org/10.14710/mdl.25.1.2025.55-63>.
37. Ding, N.; Chen, M.; Hu, L. The Design Industry in the AI Era: How AI Awareness and AI Literacy Influence the Innovative Work Behavior of Chinese Generation Y Designers. *Acta Psychol.* **2025**, *260*, 105650. <https://doi.org/10.1016/j.actpsy.2025.105650>.
38. Wu, J.; Xu, W.; Zhu, Z. From Imitation to Creation: AI Innovation Path for Architectural Design Teaching in the New Era. *Educ. Sci.* **2026**, *16*, 1078. <https://doi.org/10.3390/educsci16071078>.
39. Abrusci, L.; Dabaghi, K.; D'Urso, S.; Sciarrone, F. AI4Design: A Generative AI-Based System to Improve Creativity in Design—A Field Evaluation. *Comput. Educ. Artif. Intelligence.* **2025**, *8*, 100401. <https://doi.org/10.1016/j.caeai.2025.100401>.
40. Çiçek, S.; Aksu, M.S.; Öztürk, E.; Bingöl, K.; Mersin, G.; Koç, M.; Akmaz, O.K.; Başarır, L. ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique. *J. Comput. Design.* **2025**, *6*, 165–190. <https://doi.org/10.53710/jcode.1618548>.
41. Hikmet, Ş.; Ozay, N. Human–AI Collaboration in Architectural Design Education: Towards a Conceptual Framework. *Buildings* **2026**, *16*, 1097. <https://doi.org/10.3390/buildings16061097>.
42. Jin, S.; Tu, H.; Li, J.; Fang, Y.; Qu, Z.; Xu, F.; Liu, K.; Lin, Y. Enhancing Architectural Education through Artificial Intelligence: A Case Study of an AI-Assisted Architectural Programming and Design Course. *Buildings* **2024**, *14*, 1613. <https://doi.org/10.3390/buildings14061613>.
43. Liu, K.; Yigitcanlar, T.; Browne, W.; Fu, Y. Prompts for Planning-AI Integration: LLM Prompt Design for Supporting Sustainable Urban Development. *J. Open Innov. Technol. Mark. Complex.* **2025**, *11*, 100666. <https://doi.org/10.1016/j.joitmc.2025.100666>.
44. Lobato Valdespino, J.C.; Flores Romero, J.H. Inteligencia Expandida: Exploraciones Pedagógicas de Diseño Discursivo Texto-Imagen. In *Proceedings of the XII Jornadas Sobre Innovación Docente en Arquitectura (JIDA'24), Madrid, Spain, 21–22 November 2024*; Universitat Politècnica de Catalunya, Iniciativa Digital Politècnica: Barcelona, Spain, 2024. <https://doi.org/10.5821/jida.2024.13310>.
45. Medel-Vera, C.; Britton, S.; Gates, W.F. An Exploration of the Role of Generative AI in Fostering Creativity in Architectural Learning Environments. *Comput. Educ. Artif. Intell.* **2025**, *9*, 100501. <https://doi.org/10.1016/j.caeai.2025.100501>.
46. Obeidat, B. Generative AI in a Contemporary Architecture History and Theory Course: A Case Study of Architecture Students' Perceptions of AI-Supported Presentation Work. *Soc. Sci. Humanit. Open.* **2026**, *13*, 102951. <https://doi.org/10.1016/j.ssaho.2026.102951>.
47. Poturak, S. Integration of Artificial Intelligence in Professional Architectural Practice. *Limes-Plus* **2024**, *20*, 73–97. <https://doi.org/10.69899/limes-plus-24212-3073p>.
48. Sopher, H.; Varinlioglu, G. A Comparative Study of Two AI-Assisted Architectural Design Studios. In *Confluence-Proceedings of the 43rd Conference on Education and Research in Computer Aided Architectural Design in Europe, eCAADe, Ankara, Turkey, 1–5 September 2025*; Volume 1, pp. 873–880. <https://doi.org/10.52842/conf.ecaade.2025.1.873>.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.