Kristen M. Edwards[1]
Department of Mechanical Engineering,
Massachusetts Institute of Technology,
Cambridge, MA
email: kme@mit.edu

Farnaz Tehranchi
School of Engineering Design and Innovation,
The Pennsylvania State University,
University Park, PA
email: farnaz@psu.edu

Scarlett Miller
Industrial and Manufacturing Engineering,
The Pennsylvania State University,
University Park, PA
email: shm13@psu.edu

Faez Ahmed
Department of Mechanical Engineering,
Massachusetts Institute of Technology,
Cambridge, MA
email: faez@mit.edu

AI Judges in Design: Toward Expert-Equivalent Design Evaluations with Vision-Language Models and In-Context Learning

The subjective evaluation of early-stage engineering designs, such as concept sketches, traditionally relies on human experts. However, expert evaluations are time-consuming, expensive, and sometimes inconsistent. Recent advances in vision-language models (VLMs) offer the potential to automate design assessments, but it is crucial to ensure that these AI "judges" perform on par with human experts. This work introduces in-context learning (ICL)-enhanced VLM judges and a comprehensive statistical framework (including agreement, error, correlation, statistical difference checks, equivalence testing, and top-set overlap) to rigorously assess AI-expert equivalence. Across two case studies, we show that reasoning-enabled VLMs are the strongest-performing AI judges. They consistently outperform 2/3 trained novices across all metrics, and for measures such as uniqueness, creativity, and drawing quality, they approach expert-equivalent performance. In specific cases, they even exceed expert–expert agreement, attaining lower mean absolute error and higher rank correlations than the expert baseline. These findings suggest that, on certain statistical tests, AI judges are not only approaching expert-expert equivalence but in some cases surpassing it.

1 Introduction

In engineering design, assessing the quality and creativity of early concept sketches is a critical but challenging task [1–3]. Designers and researchers often rely on subjective evaluations by domain experts to judge the creativity, novelty, aesthetic appeal, and functional feasibility of a design concept [4–6]. This approach, while considered the gold standard, faces well-known issues: expert assessment is labor-intensive, potentially costly, and can suffer from variability due to personal biases and differing criteria. While defining expertise within a domain remains an active area of research [7], leading researchers argue that expert-level judges should possess "at least some formal training and experience in the target domain" [8]. Yet even within panels of experts, achieving consistency is non-trivial, typically necessitating multiple judges and statistical checks for inter-rater reliability (e.g., intraclass correlation coefficients) [9]. At the same time, crowdsourcing has been explored to increase throughput of design evaluations, but without careful control it may yield noisy results [10,11]. For example, studies show that averaging across a large crowd of non-experts can lead to inaccurate evaluations, with some raters providing systematically biased judgments [12]. These challenges motivate the search for scalable, reliable alternatives or complements to human expert judges.

Artificial intelligence (AI) offers a promising avenue to address this need. In particular, vision-language models (VLMs), which are AI models that combine image and text understanding, have recently achieved remarkable capabilities in interpreting visual content and producing human-like judgments [13–15]. For example, DriveLLaVA [16] is a VLM fine-tuned to make human-level judgments in driving scenarios, while Lin et al. [17] developed models capable of distinguishing AI-generated from human-created art. Their work was driven by the observation that humans can often tell the difference, but large-scale human evaluation is costly and impractical.

In engineering design, a sufficiently advanced AI could assess a sketch's creativity or functionality in seconds, enabling real-time feedback and rapid screening of large idea pools. However, deploying AI as a design "judge" demands caution: its evaluations must be as credible and valid as those of a human expert. In other words, the AI's ratings should be **equivalent to expert ratings** for us to trust its judgments in critical design decisions.

Recent developments make the goal of expert equivalence increasingly plausible. Modern VLMs and other multimodal methods have demonstrated an ability to understand drawings or images and relate them to semantic concepts. In engineering design, such models have shown high correlation with human assessments of creativity and other metrics [18–21]. However, these methods require large amounts of training data, limiting their practical use.

Preliminary work has shown that pre-trained VLMs can assess design similarity [22], and other studies explore LLMs in engineering tasks like solving mechanics problems [23] or building trusses using FEM modules [24]. These results suggest that with the right models and inputs, AI can capture the nuanced criteria used by experts.

However, a key challenge remains: rigorously verifying that AI evaluators truly match expert performance. Prior studies often report high correlations or small average errors [18,25]. While encouraging, such measures alone do not confirm that the AI's performance is indistinguishable from that of a human expert within acceptable bounds. What is needed is a rigorous statistical equivalence perspective: instead of asking whether the AI is simply correlated with or not significantly worse than humans, we ask whether any difference between the AI and human evaluations is within acceptable bounds across multidimensional criteria. Adopting statistical equivalence testing provides a more stringent and meaningful validation that the AI can serve as a surrogate for human judges with confidence and highlights areas of improvement

[1]Corresponding Author.
Version 1.18, August 7, 2026

for future AI judges.

In this paper, we address this gap by developing AI "judges" for evaluating design sketches and by introducing a comprehensive framework to test for expert equivalence. We focus on two case studies of rating early-stage design sketches on subjective metrics commonly used in conceptual design assessment (e.g., uniqueness, creativity, usefulness, drawing quality, and technical goodness). We utilize in-context learning (ICL) which is a technique by which a large-language model (LLM) or VLM learns to perform a new task by seeing a few demonstrations of that task directly in the prompt. In our case, ICL allows us to expose each AI judge to examples of design sketches and an expert's rating, thus better aligning the AI judge to the expert's preferences without retraining or changing any parameters of the base VLM. We test five ICL-enhanced VLM "judges" using a dataset of design sketches with scores obtained by multiple design experts. We then design and apply a rigorous testing procedure to determine whether the models' ratings are statistically on par with human expert ratings, within a predefined tolerance margin. To our knowledge, this work is the first to demonstrate and validate an AI model achieving expert-level equivalence in an engineering design evaluation context.

Contributions: The key contributions of this work are as follows:

- We develop an *ICL-enhanced* AI judge built on a reasoning VLM that achieves expert-equivalent performance and, in specific settings, *exceeds expert–expert agreement*. In particular, for creativity and drawing quality, we observe mean absolute error and rank-correlation levels that surpass the expert–expert baseline.

- We propose a *comprehensive statistical validation framework* to assess AI-expert agreement across multiple dimensions. Our approach integrates correlation metrics, statistical difference hypothesis testing, error metrics, equivalence testing, and top-set overlap analysis to provide a robust and quantifiable assessment of AI-expert agreement.

- We showcase two case studies rating design sketches on qualities such as uniqueness, creativity, and drawing quality, revealing that with carefully crafted in-context learning, one can coax near-expert judgments from a pre-trained VLM.
  - We show that top AI judges not only approach expert-level agreement for many metrics, but also meet expert agreement better or the same as 2/3 trained novices across all metrics.
  - We confirm this *quantitatively* by verifying equivalence with human experts using multipronged statistical checks.

- We show that *a handful of ratings* from experts can be scaled to *rate thousands of designs* using in-context learning for VLMs while maintaining statistical equivalence to experts at a level higher than the average trained novice.

The remainder of the paper is organized as follows. Section 2 (Related Works) reviews established metrics for design evaluation, the use of vision-language models in subjective evaluation tasks, in-context learning as a method of enhancing VLM performance, and statistical methods for comparing AI with humans. Sections 3 and 4 present our methodology, including the development of AI judges, the experimental setup for two case studies, and the statistical framework for evaluating their performance relative to human experts. Section 5 (Results) reports the experimental results, including quantitative comparisons and illustrative figures. Section 6 (Discussion) interprets the findings, while Section 7 outlines Limitations and Future Work. Finally, Section 8 (Conclusion) summarizes the contributions and significance of this study for the design engineering community.

2 Related Works

In the following sections, we describe related works on established metrics for evaluating engineering designs, the use of vision-language models for subjective evaluation, how in-context learning can be used to enhance VLMs even in data-scarce domains, and finally, established statistical methods for comparing raters.

2.1 Engineering Design Metrics for Design Evaluation. Engineering design literature has established a variety of metrics to evaluate the outcomes of conceptual design and ideation. A foundational distinction is that creative design quality is multi-dimensional: a concept must be both novel and functional to be truly valuable.

Various metrics are used to measure the creativity of a design or idea. These metrics include, but are not limited to, expert panels [5,26–29], the Consensual Assessment Technique (CAT) which is based on using several expert raters [8,30,31], the Shah, Vargas-Hernandez and Smith (SVS) method which outlines numerical methods to assess the novelty, variety, quality, and quantity of designs during idea generation [32], and the Comparative Creativity Assessment (CCA), which is based on the SVS method and calculates the creativity of an individual concept in a set of designs [33]. Among all the metrics created, CAT [8,30,31] is often cited as the "gold standard" of creativity metrics; however, obtaining CAT ratings is very resource intensive [7,34,35] as expert raters must code hundreds or thousands of ideas.

Many design studies employ expert judges to rate designs on Likert scales corresponding to specific design attributes. CAT uses human experts, averaging their scores to arrive at a reliable ground truth for creativity. CAT relies heavily on the intraclass correlation coefficient (ICC), and, particularly on having a sufficient ICC between experts [36]. ICC signifies "the extent to which the mean rating assigned by a group of judges is reliable"[36], and typically ICC values of 0.7 or greater are considered acceptable levels of agreement among judges [37,38]. Although experts can be consistent, they are costly to obtain. Work by Miller et al., explores the use of trained human novices as proxies for experts [9]. They find that trained novices provide design ratings that are consistent with experts for novelty but less consistent for quality, and that novice agreement with experts varies if the metrics are social science or engineering metrics.

Furthermore, algorithmic metrics (like Shah's approach for novelty or variety [32]) have been proposed, but they sometimes fail to capture the subjective nuances that experts weigh. Data-driven approaches combine features from sketches with machine learning to predict CAT-style scores [18,19], bridging the gap between purely automated metrics and human judgment. Our work builds on these approaches, leveraging machine learning to replicate expert-labeled design metrics.

2.2 Vision-Language Models in Subjective Evaluation and LLM-as-a-Judge. VLMs have shown strong performance on tasks where images and text interplay, such as image captioning. Prior design-centered work [18–21,39] often *trains* a model on a large dataset of sketches labeled with, say, creativity scores, so that the model can predict the scores for new sketches. Our approach differs: we do not train or fine-tune the underlying large model; we rely instead on in-context learning, specifically few-shot prompting [40], in which we provide the VLM with a handful of demonstration examples that illustrate how a human might label a design. The model then infers how to rate a new sketch in a similar manner. This approach is attractive in design contexts where data is scarce, or experts are reluctant to label hundreds of examples.

Recently, the concept of LLM-as-a-judge has been introduced, which uses LLMs as judges for subjective tasks [41]. Works have explored fine-tuning LLM-as-a-judge models to either enhance judgments of small language models [42] or to enhance performance for regression tasks [43]. Other work [25] investigates whether LLMs can assess chatbot responses as accurately

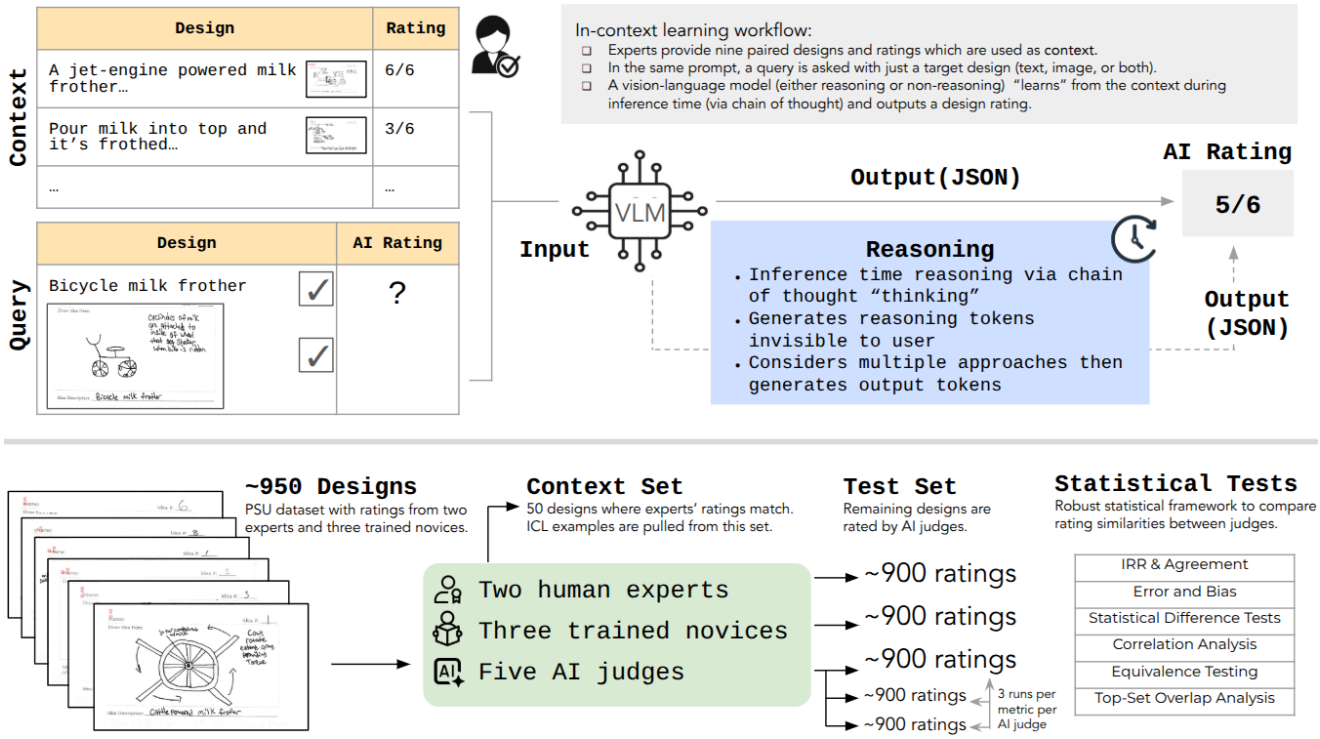


Fig. 1 Top: the in-context learning (ICL) workflow utilized to develop our five AI judges. Bottom: the experimental setup of Case Study 1. The AI judges utilize the ICL-workflow shown above to rate the test set of designs.

as humans in Chatbot Arena [44], comparing their evaluations to MT-Bench human ratings [45]. Results suggest strong LLMs (e.g., GPT-4) can approximate human judgments with 80% agreement, offering a scalable alternative to human evaluation. However, biases such as position bias (favoring the first response) and verbosity bias (favoring longer responses) remain concerns [25], which could similarly affect design evaluations. Most often, LLM-as-a-judge research only reports one statistic (agreement) when comparing LLM's ratings to those of humans, some research has also reported agreement corrected for chance or Spearman's correlation [41]. However, even with these two additional statistics, these results do not sufficiently capture if an LLM's ratings align with a human's.

In a review of VLMs for engineering design tasks, researchers tested GPT-4V's ability to assess design similarity by replicating human-based experiments [22,46]. Results show GPT-4V matches or exceeds human performance in self-consistency and transitive reasoning. Motivated by the opportunities of using VLMs as a judge, as well as the need to ensure that VLMs' ratings can match those of experts, our work explores the effectiveness of VLMs in subjective evaluation of designs. However, we move past just off-the-shelf VLMs and incorporate advancements in in-context learning to better align an AI judge with experts.

2.3 In-Context Learning. Few-shot in-context learning (ICL) is the ability of an LLM to perform a new task given a prompt that includes a few demonstrations of that task. The LLM, having seen correct input-output pairs in the demonstrations, can effectively produce a new output given a new input [40,47,48].

Thus, ICL is a way to adapt a language model to a new downstream task without fine-tuning or retraining a model, as was the traditional way of making a model work for a new task. A visual of how ICL is utilized for the AI judges in our case-study is shown in Figure 1. ICL is particularly useful for data-scarce tasks. One does not need as much data for a new task as they would need if they were fine-tuning a model, since ICL only requires a few input-output pair demonstrations [47,48].

Luo et al. identify several factors affecting ICL performance [47]: the number and format of demonstrations, their order (with later examples possibly having a greater impact), and their diversity. Generally, more demonstrations benefit LLMs, although improvements diminish as the count grows, and the maximum context size remains a constraint.

2.4 Statistical Dimensions in Prior Work. We frame AI-human alignment using six complementary dimensions established in the literature. While inter-rater agreement is the most common measure of reliability, agreement alone does not establish equivalence. We therefore consider a broader suite of metrics that capture error, bias, relative ordering, and design-selection outcomes. This ensures that where one measure is limited, another can reveal meaningful divergence or convergence between AI and human raters.

(1) Inter-rater reliability and agreement. Chance-corrected and absolute agreement are standard for ordinal or continuous ratings, commonly via Cohen's Kappa (weighted for ordinal scales) and intraclass correlation (ICC) [49–51]. In design evaluation specifically, ICC underlies the Consensual Assessment Technique, a widely accepted expert-based method for assessing creativity [9,36].

(2) Error and bias. Deviation metrics (e.g., MAE, RMSE) summarize rating error; Bland-Altman analyses visualize bias and limits of agreement beyond correlation [52,53]. On bounded Likert scales (1-6), RMSE adds little beyond MAE, but Bland-Altman remains valuable for detecting systematic over- or underestimation by models.

(3) Difference tests among groups. Nonparametric procedures such as the Friedman test with post-hoc Wilcoxon signed-rank comparisons detect overall and pairwise rating differences without assuming normality [54,55]. These tests determine whether AI-expert variability lies within the typical range of human-human variability.

(4) Correlation analysis. Spearman's ρ captures monotonic rank agreement suited to ordinal judgments [56]. High AI-expert

correlation suggests that even if absolute scores differ, the model preserves expert-like prioritization of higher- and lower-quality designs.

(5) Equivalence tests. The Two One-Sided Tests (TOST) procedure directly tests whether differences fall within a predefined margin, providing stronger evidence of practical parity than non-significant difference tests alone [57].

(6) Top-set overlap. Jaccard similarity at varying cutoffs assesses whether AI and experts select the same highest-rated items, which is critical when design workflows emphasize only the best-performing concepts [58].

3 ICL-enhanced VLM Methodology

We developed five ICL approaches for rating design sketches. Unlike typical supervised learning that requires a large training set, we show each model only a *few* demonstration examples. These include (1) a reference design sketch image, (2) a short textual description or prompt, and (3) the expert's known rating for that design. The model thus observes several (e.g., 9) such examples in the prompt (this is the context), followed by a new target design (this is the query) and must generate a rating. No parameter updates occur; the large model remains unchanged. We simply rely on the model's ability, via few-shot ICL, to mimic the rating process it saw in the examples.

3.1 Overview of our AI Judges. As described above, and shown in Figure 1, different types of information can be provided in both the context and the query. The value of using an AI judge is that the evaluation is more automated, thus it was critical that all of the information used in the context and query could be automatically retrieved. As such, we first used a vision-language model to extract the handwritten text descriptions that were on each sketch. Specifically, we instructed GPT-4o to extract the text descriptions from each design sketch via the API, and saved these descriptions as to only run this automated step once. This provided the textual description used in our later experiments.

The five AI judges that we developed each receive a description of the design task, the design metric of interest (e.g., uniqueness), and the rating scale on which to provide the ratings. Unless otherwise stated, the VLM used was GPT-4o, specifically the "gpt-4o-2024-08-06" release. The AI judges' different contexts and queries are as follows:

(1) **AI Judge: No Context**
   (a) Context: None.
   (b) Query: An image of the target design sketch.

(2) **AI Judge: Text**
   (a) Context: Nine textual descriptions of different designs and their expert ratings.
   (b) Query: A textual description of the target design.

(3) **AI Judge: Text + Image**
   (a) Context: Nine textual descriptions of different designs and their expert ratings.
   (b) Query: A textual description and an image of the target design sketch.

(4) **AI Judge: Text + Image + Reasoning**
   (a) Context: Nine textual descriptions of different designs and their expert ratings.
   (b) Query: A textual description and an image of the target design sketch.
   (c) Model: OpenAI o1 reasoning model.

(5) **AI Judge: Text + Image + o3 Reasoning (abbreviated to AI Judge: o3 Reasoning)**
   (a) Context: Nine textual descriptions of different designs and their expert ratings.
   (b) Query: A textual description and an image of the target design sketch.
   (c) Model: OpenAI o3 reasoning model.

3.2 Context Size and Selection. We ran a series of initial experiments using 50 designs to determine the standard ICL and prompting methodology used in the remainder of the study. We varied the wording of the prompt, ultimately deciding to (1) require that the model first describe the target design, then provide a justification for the rating it would assign that design, and finally provide the numerical rating, and (2) restrict the model output to a JSON format.

We also experimented with the number of context examples provided, evaluating performance for n =1, 5, 7, 9, 15, and 20 context examples. We empirically found that using 9 context examples and one target design (10 designs total) yielded the best results, and increasing to 15 or 20 context examples did not improve performance. We note that these were small-scale initial experiments done to help our team create a standard starting point from which to begin the main research.

Finally, we experimented with the modality (text and/or image) of the context examples provided. We found that including an image of the target design improved performance compared to using only a textual description. However, including images of the context examples did not improve performance; textual descriptions of the example designs were sufficient. Although this result was counterintuitive, similar findings have been reported in LLM4CAD [59] and SpatialRGPT [60], where visual inputs did not always enhance VLM performance relative to textual inputs.

Through our initial experiments exploring the number and nature of context demonstrations, we found that using nine context designs along with a tenth target query provided a reliable framework. To minimize confounding variables, we kept this constant across all judges and runs that used ICL. While systematically varying these values is an important avenue for future research, it falls outside the scope of this work. Details of the context set and selection strategy are outlined for each case study in Section 4.1.

4 Methodology for Comparing AI Judges and Human Experts in Design Ratings

We evaluated the five AI judges (AI Judge: No Context, AI Judge: Text, AI Judge: Text + Image, and AI Judge: Text + Image + Reasoning) against the ratings given by experts for each design metric. We compare each AI judge against each expert individually rather than against the averaged expert ratings to ensure that all comparisons are made against ratings that were actually assigned. Across the two case studies and all design metrics, there is an average MAE between the two experts of 1.08. Averaging ratings would therefore frequently yield values that neither expert provided. Moreover, because the Likert scale is ordinal and discrete (1-6), averaging produces continuous values (e.g., 3.5) that an AI judge constrained to the same ordinal scale would never output. Treating experts individually also enables direct comparison between AI-expert agreement and expert-expert agreement, which serves as an empirical baseline for grounding the results of statistical tests.

In Case Study 1, two human experts (Expert 1 and Expert 2) and 3 trained novices (Trained Novice 1-3) rated a total of 934 early design sketches of milk frothers. This dataset[2] originates from prior studies [61–63], where each sketch was evaluated on a 6-point Likert scale across four design metrics: *uniqueness*, *creativity*, *usefulness*, and *drawing quality*. These ratings were provided by two experts and three trained novices.

In Case Study 2, we utilized a dataset of 143 early-stage design sketches of smartphone holders rated on *uniqueness* and *technical goodness* by three experts [64]. Similarly, each design was rated on a 6-point Likert scale. In contrast to Case Study 1, in Case Study 2 there are no ratings from trained novices. The dataset can be found here[3].

[2] https://sites.psu.edu/creativitymetrics/2018/07/18/milkfrother/
[3] https://sites.psu.edu/creativitymetrics/2020/05/19/319/

In both case studies, the expert raters were classified as such based on their graduate degrees or completion of graduate coursework in an engineering design-related field [9]. The novices are trained using the modified Q-sort technique [65] in which experts select "exemplars or benchmarks of what constitutes, for example, a highly creative product and a highly uncreative product." [9]. The novices are then trained using these examples. Our overarching goals were to determine:

(1) Which AI judge most closely aligns with the human experts' ratings, and
(2) Whether any AI judges perform at a level comparable to a *human expert*, and
(3) In Case Study 1: whether any AI judges perform at a level comparable to a *trained novice*.

4.1 Data Preparation. In Case Study 1, we reserved 50 sketches, where both experts provided matching ratings, as our context set for our ICL-enhanced AI judges. The full set of 934 sketches, including these context examples, was rated *once* by two experts and three trained novices. To ensure an unbiased evaluation, only the remaining 884 sketches were included in the test set. Finally, any sketches missing ratings from any judge (expert, trained novice, or AI) were removed, resulting in a final test set of approximately 875 designs.

For each design metric, we selected nine context designs, which were provided to all AI judges except for the "No Context" judge. In each of the three runs, and independently for each design metric, these nine designs were drawn from the reserved context set to achieve two criteria: (1) agreement between the two experts, and (2) a uniform distribution across the six-point Likert scale.

In Case Study 2, we analyzed a smaller dataset of 143 designs with three expert raters. To maintain consistency with Case Study 1, we used two experts as our expert-expert baseline. We selected the pair of experts with the highest raw agreement in their ratings for our two design metrics of interest: uniqueness and technical goodness.

To address the limited dataset size, we treated each run as a separate experimental instantiation of the AI judge, corresponding to a distinct set of sampled context examples. For each run, the test set consisted of the 134 remaining designs once the nine context designs were excluded. Context designs were again chosen to satisfy the same two criteria as in Case Study 1: expert agreement and uniform coverage of the rating scale.

4.2 Validation Metrics and Statistical Analyses. To thoroughly assess each AI model's performance relative to the human experts, we combined multiple statistical approaches. These were selected based on the *nature of the data* (ordinal Likert ratings, limited to discrete integer values from 1 to 6), the *goal of measuring both absolute and relative agreement*, and the *need for nonparametric methods* given non-normal distributions. All analyses were repeated separately for each of the design metrics.

Inter-Rater Reliability and Agreement. We assess AI-human consistency using Weighted Cohen's Kappa (quadratic) and Intraclass Correlation Coefficient (ICC). Kappa accounts for chance-corrected ordinal agreement, while ICC measures absolute agreement. Higher values indicate stronger reliability.

Rationale: Kappa is suited for ordinal Likert data, adjusting for partial agreement, while ICC treats ratings as continuous and reflects both correlation and agreement. Together, they ensure a robust assessment of AI performance relative to human experts.

Error and Bias. We measure error and bias using Mean Absolute Error (MAE) and Bland-Altman analysis of systematic bias. MAE quantifies rating deviations:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{r}_i - r_i| \tag{1}$$

where $\hat{r}_i$ is the AI rating and r_i is the expert rating. AI models are considered human-level if their MAE matches or is lower than expert-expert MAE.

Bland-Altman analysis detects systematic bias by measuring mean rating differences and their variability. This demonstrates both the mean difference (or bias) between raters, as well as the standard deviation (SD) between the raters. If the bias is close to zero and most data points fall within ±1.96× SD of that difference, the two raters exhibit strong agreement. If AI-expert differences fall within the expert-expert limits of agreement (LoA), the AI demonstrates human-like rating consistency.

Rationale: MAE directly measures rating accuracy and allows comparison against expert disagreement levels. Since there is no absolute ground truth, we compute MAE separately for each expert reference.

Standard correlation does not always reveal systematic rating biases. Bland-Altman analysis highlights potential biases and the spread of disagreements at different rating levels, thereby indicating if the AI's errors are consistent with normal human disagreement. Bland-Altman plots can reveal any systematic trends (heteroscedasticity). For example, perhaps the models agree with humans on low scores but diverge on very high scores. A trend line in the Bland-Altman (differences growing with the mean rating) would indicate the agreement depends on the sketch's overall quality rating.

Statistical Difference Tests. We use the **Friedman test** to detect overall rating differences between experts, AI models, and trained novices. If significant ($p < 0.05$), **Wilcoxon signed-rank tests** with Bonferroni correction identify specific AI-human differences.

Rationale: The Friedman test is non-parametric and suitable for repeated measures. Pairwise Wilcoxon tests pinpoint which AI models differ significantly from experts while controlling for multiple comparisons.

Correlation Analysis. Spearman's rank correlation measures whether AI models maintain the relative ordering of sketches. A high AI-expert correlation near the expert-expert correlation suggests strong alignment in ranking.

Rationale: Even if absolute ratings differ, AI models should ideally rank sketches similarly to experts. Spearman's ρ evaluates this ranking agreement.

Equivalence Testing. The **Two One-Sided Tests (TOST)** framework assesses whether AI ratings are statistically equivalent to expert ratings within a predefined tolerance (e.g., ±1.0 points). A significant result confirms AI-expert rating similarity.

Rationale: Non-significant differences (e.g., from Friedman or Wilcoxon tests) do not confirm equivalence; they only show *no evidence of difference*. TOST directly tests whether any differences are *small enough* to be practically irrelevant, reinforcing or contradicting findings of "no difference."

Equivalence testing complements hypothesis testing. A non-significant difference test means no detected difference, but a non-significant equivalence test means insufficient evidence of equivalence—often due to variability. If both tests support equivalence, there is good evidence that an AI judge is matching an expert. If difference is non-significant but equivalence fails, the AI is close but lacks definitive proof of expert-level performance.

Top-Set Overlap Analysis. Another practical way to evaluate whether the AI models identify the same "best" sketches as experts is to measure *set overlap* in the top-rated designs. Although correlation and MAE capture broad patterns, some design use cases focus primarily on whether a model picks out the same top fraction of designs that a human would. We can measure this by looking at Jaccard Similarity at variable top-set cutoffs.

To compare an AI model and an expert at multiple thresholds, we define the following methodology:

(a) **Nominal Fraction**. We start by choosing a series of nominal cutoffs (e.g., 5%, 10%, 15%, ...) to indicate the intended top k% of items for that expert's ratings.

(b) **Reference Expert Top Set + Ties**. For each fraction f, let $N = \lceil f \cdot (\text{total sketches}) \rceil$. We collect the top N sketches *from the expert*, *including all ties* at the boundary rating. In practice, if $N = 10$ but the 10th and 11th sketches share the same rating, we include both. As a result, the final set may exceed N items, yielding an *actual fraction* larger than f.

(c) **Model Top Set + Ties (Matching Size)**. Suppose the expert's top set ends up being N_e sketches (after ties). We then form the model's top set by taking *at least* N_e items under the same tie rule. This ensures neither the expert set nor the model set excludes items arbitrarily cut off by rating ties.

(d) **Jaccard Similarity**. Denote the expert's top set by E_{top} and the model's top set by M_{top}. We measure

$$\text{Jaccard}(E_{\text{top}},\ M_{\text{top}}) \ = \ \frac{\left|E_{\text{top}} \ \cap \ M_{\text{top}}\right|}{\left|E_{\text{top}} \ \cup \ M_{\text{top}}\right|}.$$

A higher Jaccard indicates better agreement on which sketches are considered best.

(e) **Plotting Actual Fraction vs. Jaccard**. To compare the models and experts, we record the *actual* fraction of total sketches $|E_{\text{top}}|/(\text{total sketches})$ on the x-axis, since boundary ties can inflate (or occasionally reduce) the set size. On the y-axis, we plot the Jaccard similarity for each fraction. Doing this for multiple fractions (5%, 10%, 15%, etc.) produces a curve showing how well the model and the expert align on which items belong in progressively larger top sets.

(f) **Comparing AUC of the Jaccard Top-Set Similarity** To quantitatively measure this metric, we find the AUC for the Jaccard plots for Expert 1 vs. Expert 2, and for all five AI judges vs. each expert as well as the three trained novices vs. each expert. Figure 3 shows an example of this plot and the AUC scores are reported in each of the run tables.

Rationale: Whereas correlation tests broad item-by-item rank alignment, top-set overlap directly answers, "Does the model pick out the same top-tier sketches that the expert considers best?" This is often crucial in design or idea-generation tasks where only the highest-performing concepts matter. By including ties at the threshold, we avoid arbitrary cutoffs (e.g., one design that shares the boundary rating gets left out of the model set but included in the expert set). Furthermore, comparing multiple fractions reveals if a model aligns strongly at small cutoffs (the absolute top 5%) but diverges beyond that, or vice versa.

4.3 Decision Rules for "Human-Level" Ratings.

General thresholding. Our goal is to assess how well AI judges can learn and approximate expert ratings, and we assess this across a diverse set of agreement metrics with heterogeneous scales, including bounded measures (e.g., Kappa, ICC, Spearman, Jaccard AUC), unbounded error-based measures (e.g., MAE), and hypothesis-test outcomes. To enable meaningful interpretation of the scores, we anchor all thresholds to expert-expert performance.

We deem an AI judge "as good as" a human expert for a given metric if the AI-expert score falls within 20% of the corresponding expert-expert score, in the direction of worse performance. Formally, let M be a positively oriented metric ("higher is better": Kappa, ICC, Spearman, Jaccard AUC). The AI-expert pair *passes* if

$$M_{\text{AI, expert}} \ \geq \ 0.8 \times M_{\text{expert, expert}}.$$

For negatively oriented metrics ("lower is better": MAE, |bias|, LoA width), the AI-expert pair passes if

$$M_{\text{AI, expert}} \ \leq \ 1.2 \times M_{\text{expert, expert}}.$$

This relative threshold is intentionally defined with respect to expert-expert performance rather than absolute metric values, reflecting the fact that expert agreement is itself imperfect and varies across metrics. The 20% margin introduces controlled lenience that accounts for finite-sample variability, metric sensitivity, and non-zero expert disagreement, while still requiring AI performance to remain close to the expert baseline.

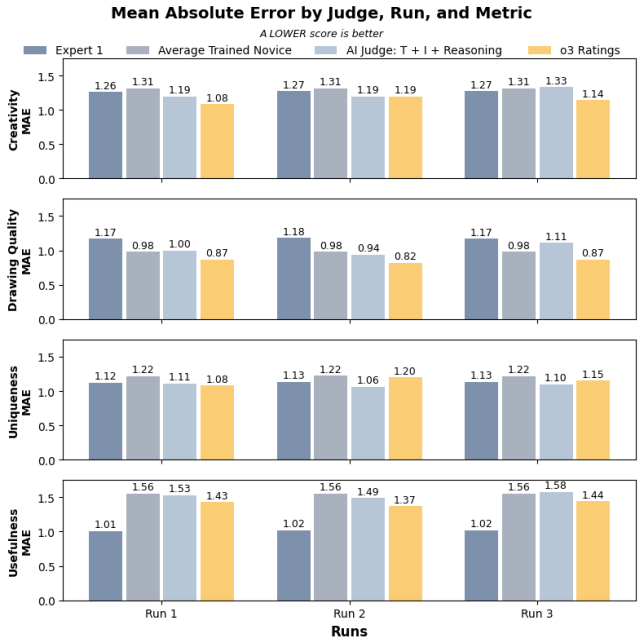
Furthermore, we observe that most AI judges agree with an expert less closely than two experts agree with each other. As a result, a strict cutoff at the expert-expert score would cause most AI judges to fail across nearly all tests, obscuring meaningful differences in relative performance among AI judges and reducing the interpretability of the results. The exact threshold percentage could be adjusted in future applications depending on the desired strictness of expert-level equivalence. Overall, anchoring the threshold to expert-expert agreement ensures that all conclusions remain grounded in human performance.

Exceptions. (i) **Equivalence:** We use an equivalence margin of ±1.0 in our TOST setup. This choice is motivated by the observed disagreement between expert raters: across metrics and case studies, the MAE between expert ratings ranged from 0.76 to 1.27, with an average of 1.08. Thus, a one-point difference on the 1–6 Likert scale reflects typical expert-level variation. Using an integer-valued margin is also consistent with the discrete nature of the Likert rating scale and avoids introducing artificial precision beyond the resolution of the data.

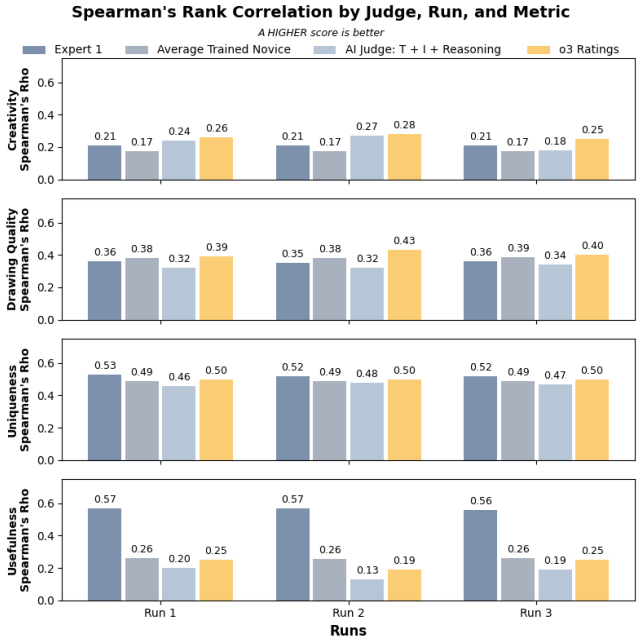
(ii) **Difference tests:** If the Friedman test is significant ($p > 0.05$), we move on to the Wilcoxon signed-rank test. The Bonferroni-corrected Wilcoxon p-value must be non-significant ($p > 0.05$) for the AI vs. expert to pass this test.

We repeated these analyses independently for each of the design metrics in each case study, so uniqueness, creativity, usefulness, and drawing quality for Case Study 1 and uniqueness and technical goodness for Case Study 2. We then summarized which judge(s) met which criteria, and thereby concluded which judge performed better and whether it operated at a "expert-level" for any or all of the metrics. It is possible an AI judge is very good on some metrics but not others, or beyond a trained-novice level but not yet at an expert level. We note such cases in the Results section and discuss overarching trends in the Discussion.

Practical Interpretation: If an AI consistently demonstrates close agreement metrics, MAE ≃ expert-expert MAE, low bias and similar limits of agreement as expert-expert pairs, a TOST equivalence outcome, no significant difference in distribution, and a strong correlation with the expert, then it can be treated as being "as reliable as" a human rater on that metric. Otherwise, we can examine where it fell short (e.g., higher error, lower correlation, or failing to show equivalence) to pinpoint which aspects of performance need improvement. Lastly, we compared the five AI judges in two respects: 1) among each other to determine which ICL setup yielded ratings that best matched experts, and 2) against the three trained human novices (in Case Study 1) on the same criteria to identify if a well-developed AI judge shown just nine design-rating examples could equally or more closely match expert ratings than trained humans.



(*a*) Mean absolute error across judges. Lower is better.



(*b*) Spearman correlation across judges. Higher is better.

Fig. 2 Case Study 1: The AI-judge which utilizes o3 Reasoning, shown in yellow, outperforms the average trained novice in MAE for all runs across all design metrics. Perhaps more significantly, it also outperforms the upper baselines (expert-expert MAE and Spearman's rank correlation) for all creativity and drawing quality runs.

5 Results

In this section, we report our findings for the two case studies. We compare the ratings from the five AI judges (No Context, Text, Text + Image, and Text + Image + o1 Reasoning, and Text + Image + o3 Reasoning (abbreviated as o3 Reasoning)) to the expert raters in each case study, and report their statistical results separately for each design metric (e.g., uniqueness).

How to Read the Results We show two example tables of the entire set of statistical results from one run for one design metric (Tables 1 and 6). In these tables, we report the baseline expert-expert results of the statistical tests outlined in Section 8, as well as the results from each of the other human judges and AI judges compared to Expert 1 and then compared to Expert 2. If these results meet the thresholds described in Section 4.3 we display them in **bold** and count them toward the number of tests passed. The total number of tests is 9, corresponding to each of the columns, with the Bland-Altman limits of agreement comprising two tests: the mean difference *and* the standard deviation of the difference. We include one plot of the Jaccard similarity at various fractional cutoffs (Figure 3) to compare the various judges and help visualize the Jaccard area under the curve (AUC) reported in the statistical results. Since these comprehensive tables are quite dense, the rest are provided in the supplementary material.

For each design metric in each case study, we also provide summary results from all three runs (Tables 2- 5 for Case Study 1 and Tables 7- 8). The summary results are intended to allow for easier comparison of different judges. The number of tests passed (out of nine) for each run is shown as well as the average across all three runs. A higher average is better. In these tables, the results from the non-AI judges are shown with a gray background.

5.1 Case Study 1. For Case Study 1, we compare the ratings from the five AI judges (No Context, Text, Text + Image, and Text + Image + o1 Reasoning, and o3 Reasoning) against the ratings from the two human experts (Expert 1 and Expert 2). The expert-expert scores for each statistical test serve as a baseline. For each metric, we provide the summarized results from all three runs (Tables 2, 3, 4, and 5). For clarity for the reader, we also provide one example of the comprehensive statistical results from a single run for the creativity metric in Table 1. The comprehensive statistical results for all other runs are provided as supplementary material.

We report statistical measures of inter-rater reliability and agreement, error, statistical hypothesis tests for differences among groups, correlation, equivalence, and top-set overlap. The nine statistical tests that we report in Tables 1-8 are outlined in Section 4.2.

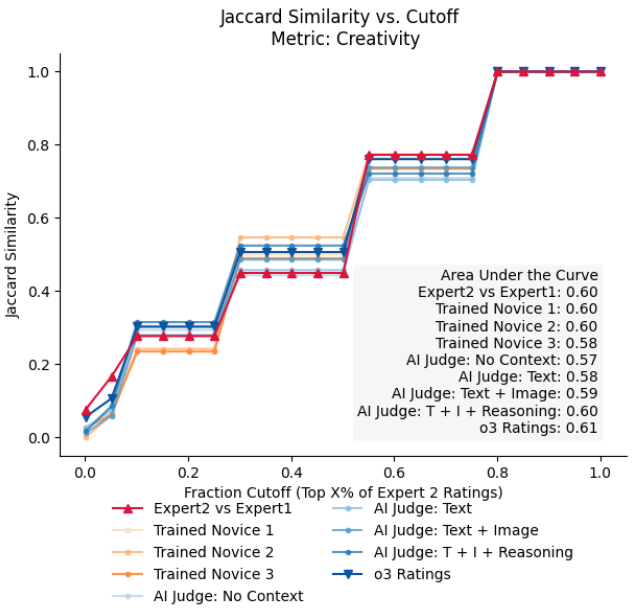


Fig. 3 The AI Judge: Text + Image + o3 Reasoning (abbreviated as o3 Ratings in the legend) reaches a higher AUC than all of the trained novices and the expert-expert agreement, therefore exceeding the upper baseline.

Table 1 Run 1: Summary of Creativity results. Kappa and ICC are relative to the expert-expert baseline of 0.24. MAE values compare to the observed Expert-Expert MAE = 1.26. Mean Diff ± 1.96 SD reports the 95% limits of agreement from a Bland-Altman analysis; this reflects the spread of itemwise differences. "Equiv?" indicates whether TOST found AI-expert equivalence within ±1.0 margin. Bold indicates model is within 20% of expert-expert agreement.

	Comparison	Kappa	ICC	MAE	Mean Diff ± 1.96 SD	Equiv?	Spear.	Wilcoxon p-corr	AUC	Tests Passed
	Expert 1 vs Expert 2	0.24	0.24	1.26	0.20 ± 3.16	-	0.21	$1.19 \cdot 10^{-03}$	0.58	-
Expert 1	Trained Novice 1	-0.01	-0.01	1.62	−0.88 ± **3.57**	**True**	-0.02	$7.63 \cdot 10^{-38}$	**0.56**	3/9
	Trained Novice 2	0.09	0.09	**1.37**	0.61 ± **3.38**	**True**	0.13	$2.69 \cdot 10^{-23}$	**0.57**	4/9
	Trained Novice 3	-0.06	-0.05	**1.50**	−0.57 ± **3.60**	**True**	-0.05	$4.88 \cdot 10^{-17}$	**0.57**	4/9
	AI Judge: No Context	0.05	0.05	2.04	−1.88 ± **2.96**	False	0.11	$1.04 \cdot 10^{-113}$	**0.58**	2/9
	AI Judge: Text	0.10	0.10	**1.49**	−1.02 ± **3.01**	False	0.14	$1.51 \cdot 10^{-59}$	**0.58**	3/9
	AI Judge: Text + Image	0.10	0.10	**1.40**	−0.56 ± **3.25**	**True**	0.10	$2.57 \cdot 10^{-20}$	**0.59**	4/9
	AI Judge: T + I + Reasoning	0.14	0.14	**1.37**	−0.91 ± **2.85**	**True**	**0.20**	$2.08 \cdot 10^{-54}$	**0.59**	5/9
	AI Judge: o3 Reasoning	0.15	0.15	**1.29**	−0.57 ± **2.99**	**True**	0.16	$1.63 \cdot 10^{-25}$	**0.57**	4/9
Expert 2	Trained Novice 1	0.15	0.15	**1.33**	−0.68 ± **3.08**	**True**	**0.17**	$2.64 \cdot 10^{-31}$	**0.60**	5/9
	Trained Novice 2	0.13	0.13	**1.37**	0.81 ± **3.12**	**True**	0.14	$1.13 \cdot 10^{-40}$	**0.60**	4/9
	Trained Novice 3	0.19	0.19	**1.23**	−0.37 ± **2.98**	**True**	**0.21**	$2.52 \cdot 10^{-11}$	**0.58**	5/9
	AI Judge: No Context	0.08	0.08	1.83	−1.68 ± **2.70**	False	**0.18**	$2.24 \cdot 10^{-112}$	**0.57**	3/9
	AI Judge: Text	0.17	0.17	**1.27**	−0.82 ± **2.71**	**True**	**0.23**	$1.01 \cdot 10^{-50}$	**0.58**	5/9
	AI Judge: Text + Image	0.17	0.17	**1.18**	−0.35 ± **2.97**	**True**	**0.19**	$8.36 \cdot 10^{-12}$	**0.59**	5/9
	AI Judge: T + I + Reasoning	0.18	0.18	**1.19**	−0.71 ± **2.63**	**True**	**0.24**	$1.99 \cdot 10^{-42}$	**0.60**	5/9
	AI Judge: o3 Reasoning	**0.24**	**0.24**	**1.08**	−0.36 ± **2.67**	**True**	**0.26**	$3.41 \cdot 10^{-14}$	**0.61**	7/9

Table 2 Creativity: Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Trained Novice 1	3	3	3	3/9
	Trained Novice 2	4	4	4	4/9
	Trained Novice 3	4	4	4	4/9
	AI Judge: No Context	2	2	2	2/9
	AI Judge: Text	3	3	3	3/9
	AI Judge: Text + Img	4	4	4	4/9
	AI Judge: Txt. + Img. + Reas.	5	3	4	4/9
	AI Judge: o3 Reasoning	4	4	4	4/9
Expert 2	Trained Novice 1	5	5	5	5/9
	Trained Novice 2	4	4	4	4/9
	Trained Novice 3	5	5	5	5/9
	AI Judge: No Context	3	3	3	3/9
	AI Judge: Text	5	5	4	4.67/9
	AI Judge: Text + Img	5	4	4	4.33/9
	AI Judge: Txt. + Img. + Reas.	5	7	5	5.67/9
	AI Judge: o3 Reasoning	7	7	7	7/9

5.1.1 Top AI judges utilize reasoning. In both case studies, either of the two reasoning judges (AI Judge: Text + Image + Reasoning or o3 Reasoning) are the first or second best performing AI judges in terms of tests passed for all design metrics. This can be seen in Tables 2, 3, 4, 5, 7, and 8.

5.1.2 Exceeding expert-expert equivalence. Zooming in on one statistic at a time, Figure 2(*a*) shows that for mean absolute error (MAE) the o3 reasoning AI judge is able to outperform the expert-expert baseline for all creativity and drawing quality runs. The o3 reasoning AI judge is also able to exceed the expert-expert baseline for Spearman's rank correlation for the same two metrics, as seen in Figure 2(*b*).

Furthermore, for creativity, drawing quality, and uniqueness, the o3 reasoning AI judge passes the MAE test (scores $\leq 1.2\times$ expert-expert MAE) for all runs. This means the o3 reasoning AI judge exceeded or was near expert-expert MAE for all design metrics but usefulness.

5.1.3 Top AI judges outperform the average trained novice. The top AI judge beats or matches 2/3 trained novices in the overall number of tests passed for creativity, uniqueness, and drawing quality. For usefulness, the top AI judge beats all three trained novices when comparing against Expert 1, but only 1/3 trained novices when comparing against Expert 2 (Table 5).

5.1.4 Looking at one statistic vs. the suite. While focusing on a single statistic such as MAE or Spearman's correlation highlights standout cases where reasoning-enabled AI judges exceed expert-expert agreement, the broader statistical suite provides a more nuanced picture. For instance, when looking at MAE or Spearman's rank correlation, creativity and drawing quality are the metrics where top AI judges exceed expert-levels of equivalence and those of the average trained novice. However, when looking at the number of tests passed, the uniqueness becomes a standout metric where the top AI judges beat or match 2/3 trained novices and pass more tests on average than they do for creativity. This underscores the importance of evaluating AI judges across the full suite of metrics rather than relying on any one measure in isolation.

5.1.5 Trends among design metrics. There are specific trends that can be seen among the design metrics. For example, usefulness is the metric for which AI judges and trained novices alike perform the worst, often passing 1-3 tests out of nine (Table 5). In the following paragraphs, specific trends for each design metric are outlined.

Creativity. For the creativity metric, the strongest performance consistently comes from the o3 Reasoning judge. In Run 1 shown in Table 1, o3 Reasoning achieves the highest reliability values of all raters (Weighted Cohen's Kappa = 0.24, ICC = 0.24) when compared to Expert 2, effectively matching the expert-expert baseline. It also attains the lowest MAE (1.08, compared to the expert-expert MAE of 1.26) and passes the largest number of statistical tests (7/9), outperforming both trained novices and the other AI judges.

Across all three runs, o3 Reasoning remains the most reliable AI judge, averaging 7/9 tests passed with Expert 2, which exceeds the performance of any trained novice (Table 2). Moreover, the o3 Reasoning judge outperforms expert-expert baselines for MAE and Spearman's rank correlation for all three creativity runs, as shown in Figure 2. In Figure 3 that trend continues: the o3 Reasoning AI judge has the highest AUC in the Jaccard similarity plot, and we can observe that the o3 Reasoning line quickly catches up to the expert-expert line, and ultimately exceeds the expert-expert AUC.

The Text + Image + Reasoning judge is the next best AI performer. It reaches equivalence on MAE and often passes 5-6 tests out of 9, making it comparable to the top trained novices and sometimes stronger. By contrast, the No Context judge performs worst among the AI judges, consistently failing to reach equivalence and passing only 2-3 tests.

Interestingly, the mean difference for all but one non-expert rater is negative (as seen in Table 1 for run 1 as well as in the comprehensive results for runs 2 and 3 provided as supplementary material). This means that non-expert raters consistently give *higher* creativity scores than experts do. Looking at o3 Reasoning alone, a Kruskal-Wallis test (with Mann-Whitney U post hoc) indicates differences in median ratings between o3 and the experts.

The Kruskal-Wallis tests detected robust group differences in every run ($H = 113.9\text{-}187.6$, $p \leq 1.9 \times 10^{-25}$). Post hoc Mann-Whitney U tests (Bonferroni-corrected) show that o3 Reasoning assigns higher creativity ratings than both experts, with negative Cliff's δ of moderate magnitude relative to Expert 2 ($\delta \approx -0.18$ to -0.26) and moderate-large magnitude relative to Expert 1 ($\delta \approx -0.28$ to -0.34; all corrected $p < 10^{-15}$). The two experts also differ with a small but consistent effect (Expert 1 lower than Expert 2: $\delta \approx -0.098$, corrected $p \approx 5.6 \times 10^{-4}$). Overall, the omnibus effect is driven primarily by o3 Reasoning's systematically higher creativity scores relative to both experts, plus a smaller Expert 1 vs. Expert 2 scale shift.

Consistent with this, Cliff's δ is negative (≈ -0.35 to -0.15), which, under our sign convention (*experts − o3*), means o3 assigns higher ratings than both experts.

Taken together, these results indicate that reasoning-enabled multimodal judges (especially o3 Reasoning) not only approach, but in many cases surpass, the consistency of trained novices for creativity ratings.

Table 3 Uniqueness: Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Trained Novice 1	7	8	8	7.67/9
	Trained Novice 2	5	5	5	5/9
	Trained Novice 3	9	9	9	9/9
	AI Judge: No Context	2	2	2	2/9
	AI Judge: Text	5	5	4	4.67/9
	AI Judge: Text + Img.	5	5	6	5.33/9
	AI Judge: Txt. + Img. + Reas.	6	8	8	7.33/9
	AI Judge: o3 Reasoning	8	7	8	7.67/9
Expert 2	Trained Novice 1	9	9	9	9/9
	Trained Novice 2	3	3	3	3/9
	Trained Novice 3	8	8	8	8/9
	AI Judge: No Context	2	2	2	2/9
	AI Judge: Text	9	9	8	8.67/9
	AI Judge: Text + Img.	4	4	5	4.33/9
	AI Judge: Txt. + Img. + Reas.	9	8	9	8.67/9
	AI Judge: o3 Reasoning	7	7	7	7/9

Uniqueness. For rating uniqueness, the best-performing AI raters were AI Judge: o3 Reasoning and AI Judge: Text + Image + Reasoning. Against Expert 1, these models consistently passed between 7/9 and 8/9 statistical tests, averaging 7.67/9 and 7.33/9, respectively. Against Expert 2, o3 Reasoning was slightly weaker, averaging 7/9, while Text + Image + Reasoning tied with AI Judge: Text as the top-performing AI judges, each averaging 8.67/9 tests passed (Table 3).

Furthermore, the top AI judges beat or matched 2/3 trained novices in terms of number of tests passed (Table 3). Relative to trained novices, the best AI judges consistently outperformed Trained Novice 2, who averaged only 3/9-5/9 tests passed. They matched or slightly lagged Trained Novice 1, who averaged 7.67/9 with Expert 1 and a perfect 9/9 with Expert 2. Compared to Trained Novice 3, who was the strongest novice with averages of 9/9 (Expert 1) and 8/9 (Expert 2), the top AI judges approached but did not surpass this level of performance. The No Context judge was the weakest performer overall, passing only 2/9 tests in every run (Table 3).

Looking at Figure 2, we also note that AI judges were unable to exceed expert-expert equivalence for MAE or Spearman's rank correlation, unlike the results for creativity or drawing quality.

At the same time, Kruskal-Wallis and post hoc analyses revealed that uniqueness ratings differed systematically among raters. Expert 1 consistently gave lower uniqueness scores than Expert 2 (Cliff's $\delta \approx -0.12$, corrected $p \approx 2.6 \times 10^{-5}$). Relative to the experts, o3 Reasoning tended to give slightly lower scores than Expert 1 (small effect, $\delta \approx 0.07\text{-}0.17$) and moderately lower scores than Expert 2 ($\delta \approx 0.19\text{-}0.28$, all corrected $p < 10^{-12}$). Thus, while o3 Reasoning achieved high rates of test-passing, it gave systematically lower ratings for uniqueness compared to both experts, with closer alignment to Expert 2 than Expert 1.

Table 4 Drawing Quality: Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Trained Novice 1	5	5	5	5/9
	Trained Novice 2	2	2	2	2/9
	Trained Novice 3	5	5	5	5/9
	AI Judge: No Context	8	8	6	7.33/9
	AI Judge: Text	5	5	8	6/9
	AI Judge: Text + Img.	5	5	6	5.33/9
	AI Judge: Txt. + Img. + Reas.	8	8	8	8/9
	AI Judge: o3 Reasoning	3	5	5	4.33/9
Expert 2	Trained Novice 1	8	8	8	8/9
	Trained Novice 2	8	8	8	8/9
	Trained Novice 3	8	8	8	8/9
	AI Judge: No Context	8	8	8	8/9
	AI Judge: Text	8	5	5	6/9
	AI Judge: Text + Img.	6	6	5	5.67/9
	AI Judge: Txt. + Img. + Reas.	8	8	6	7.33/9
	AI Judge: o3 Reasoning	8	9	9	8.67/9

Drawing Quality. For drawing quality, results were quite different when comparing against ratings from Expert 1 or Expert 2. Generally, the AI Judge: No Context, AI Judge: Text + Image + Reasoning, and AI Judge: o3 Reasoning were the top-performing AI judges. Unlike in most design metrics, AI Judge: No Context performed quite well, passing an average of 7.33/9 tests against Expert 1 and 8/9 tests against Expert 2, as shown in Table 4. The strong performance of AI Judge: No Context suggests that raw visual assessment alone is highly effective for evaluating drawing quality. This finding is supported by prior findings in [18]. Meanwhile, AI Judge: Text + Image + Reasoning demonstrated consistently strong agreement with expert ratings, reinforcing the benefits of multimodal reasoning. The o3 Reasoning judge interestingly performs much better when compared to Expert 2 than Expert 1 (as seen in Table 4).

Both AI Judge: No Context and AI Judge: Text + Image + Reasoning had superior overall performance than the trained novices

when compared to ratings from Expert 1. When comparing to ratings from Expert 2, only the o3 Reasoning AI judge surpasses all of the trained novices.

Drawing quality is one of the design metrics where the best AI judges are able to meet expert-expert agreement levels consistently. Interesting, while the trained novices perform relatively poorly compared to Expert 1, they are often able to exceed expert-expert agreement with Expert 2 for the metrics of Kappa, ICC, MAE, and Spearman correlation coefficient. These results are shown in part by Figure 2 and also in the comprehensive results from runs 1-3 in the supplementary material. AI Judge: No Context, AI Judge: Text + Image + Reasoning, and AI Judge: o3 Reasoning were similarly able to better match Expert 2 than Expert 1 was based on MAE (Figure 2).

This is bolstered by the Kruskal-Wallis tests, which indicated significant group differences across all runs (H = 176.5-304.1, $p \leq 9.3 \times 10^{-67}$). Post hoc tests show a stable expert-expert scale gap (Expert 1 > Expert 2; $\delta \approx 0.30$, corrected $p \approx 5.2\text{-}5.7 \times 10^{-30}$). Relative to the experts, o3 Reasoning is lower than Expert 1 with moderate to large effects ($\delta \approx 0.30$ to 0.45, all corrected $p < 10^{-29}$ under the Expert 1 vs. o3 ordering), but broadly comparable to Expert 2: significant in run 1 with a small effect (Expert 2 > o3; $\delta \approx 0.13$, corrected $p \approx 1.7 \times 10^{-6}$) and nonsignificant after correction in runs 2-3 (small $|\delta| \leq 0.05$). Hence, the omnibus effect is driven primarily by Expert 1's higher scale and its gap to o3 Reasoning, while o3 Reasoning and Expert 2 are closely aligned.

Table 5 Usefulness: Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Trained Novice 1	2	2	2	2/9
	Trained Novice 2	1	1	1	1/9
	Trained Novice 3	3	3	4	3.33/9
	AI Judge: No Context	2	3	3	2.67/9
	AI Judge: Text	3	2	3	2.67/9
	AI Judge: Text + Img.	2	2	3	2.33/9
	AI Judge: Txt. + Img. + Reas.	3	3	2	2.67/9
	AI Judge: o3 Reasoning	4	4	3	3.67/9
Expert 2	Trained Novice 1	2	2	2	2/9
	Trained Novice 2	1	1	1	1/9
	Trained Novice 3	2	2	2	2/9
	AI Judge: No Context	1	2	2	1.67/9
	AI Judge: Text	2	1	2	1.67/9
	AI Judge: Text + Img.	1	1	2	1.33/9
	AI Judge: Txt. + Img. + Reas.	2	1	1	1.33/9
	AI Judge: o3 Reasoning	2	1	2	1.67/9

Usefulness. Usefulness is the design metric for which the experts best agree. They have the highest Weighted Cohen's Kappa, ICC, Spearman, and AUC and the lowest MAE and mean difference (shown in the supplementary material).

In part due to the fact that passing tests is based on the expert-expert baseline, trained novices and AI judges alike perform poorly for this design metric. For rating usefulness, no AI judge consistently outperformed the others. Against Expert 1, o3 Reasoning was the strongest AI, averaging 3.67/9 tests passed, slightly ahead of the other AI judges, which clustered between 2.33/9 and 2.67/9 (Table 5). Against Expert 2, however, all AI judges performed comparably poorly, averaging between 1.33/9 and 1.67/9 tests passed. Reasoning, which improved performance for other metrics, did not confer a clear advantage here.

Compared to trained novices, the AI judges were competitive but not superior. Trained Novice 3 performed best of the novices, averaging 3.33/9 tests passed with Expert 1 and 2/9 with Expert 2. Trained Novices 1 and 2 scored lowest, averaging only 2/9 and 1/9 tests, respectively, showing that usefulness is a difficult metric for both human novices and AI judges to align with experts. Furthermore, as seen in the supplementary material, the mean difference for all of the raters was negative, indicating that all of the raters give *higher* usefulness scores than the experts (similar to the result for uniqueness). Consistent with this, Kruskal-Wallis tests across all runs revealed significant group differences (H = 41.2-166.5, $p \leq 1.2 \times 10^{-9}$) between o3 Reasoning and the two experts. Post hoc Mann-Whitney U tests with Bonferroni correction confirmed that o3 Reasoning assigns higher usefulness ratings than both experts, with negative Cliff's δ values ($\delta \approx -0.30$ to -0.14). By contrast, Expert 1 vs. Expert 2 showed no difference ($p \approx 0.99$, $\delta \approx 0$), indicating that the observed effect stems from o3-expert divergence rather than expert-expert disagreement.

Taken together, these results indicate that usefulness is the hardest metric for both humans and AI to predict. All AI judges and trained novices performed below the levels observed for creativity, uniqueness, and drawing quality, and none approached expert-expert agreement.

5.2 Case Study 2. Tables 6, 7, and 8 show the results for Case Study 2. In this case study there were three experts and no trained novices. Furthermore, the design metrics of interest were uniqueness and technical goodness. Technical goodness was defined in the study that created this dataset as a metric that "assesses the level to which each solution suits the AM processes, both in terms of capabilities and limitations" [64].

5.2.1 Which AI judges perform best. Across both metrics, the strongest performing AI judges were those that incorporated textual or multimodal inputs, while the No Context judge consistently underperformed. For uniqueness, AI Judge: Text, AI Judge: Text + Image, and AI Judge: o3 Reasoning frequently achieved near-expert levels of agreement, particularly when compared against Expert 2 (Table 7). For technical goodness, performance was more modest overall, but o3 Reasoning stood out as the best AI judge, reaching an average of 7.33/9 tests passed against Expert 2 (Table 8). By contrast, the Text + Image + Reasoning judge performed relatively poorly for technical goodness, averaging only 2-3 tests passed across experts.

5.2.2 Performance differs compared to different experts. The results show that performance varied substantially depending on which expert was used as the reference. For uniqueness, nearly all AI judges performed much better when compared to Expert 2 than when compared to Expert 1. For example, o3 Reasoning achieved 8.33/9 against Expert 2, compared to 3.33/9 against Expert 1. Even Expert 3 performed better when compared to Expert 2 than Expert 1. This suggests that Expert 2's ratings were more closely aligned with the models' outputs, whereas Expert 1's ratings diverged more. A similar though less pronounced pattern appeared for technical goodness, where o3 Reasoning again showed stronger alignment with Expert 2 than with Expert 1. These findings highlight the sensitivity of AI-expert equivalence to the specific expert chosen for comparison.

For uniqueness, this aligns with results from Kruskal-Wallis tests, which revealed significant group differences across all runs (H = 34.3-53.6, $p < 10^{-7}$). Post hoc Mann-Whitney U tests show that Expert 1 consistently diverges from both Expert 2 and Expert 3, with moderate effect sizes (Cliff's $\delta \approx 0.32\text{-}0.37$). By contrast, Expert 2 and Expert 3 did not differ significantly ($p \approx 0.67\text{-}0.98$, $\delta \approx 0$).

Relative to the experts, the o3 Reasoning judge assigns systematically higher uniqueness scores than Expert 1, with large effect sizes (Cliff's $\delta \approx 0.33\text{-}0.50$, all corrected $p < 10^{-11}$). Against Expert 2 or Expert 3, however, o3's ratings were not significantly different (Bonferroni-corrected $p \geq 0.18$, $\delta \approx -0.03\text{-}0.15$).

These results indicate that Expert 1 applies a more conservative scale for uniqueness, leading to systematic differences both with

Table 6 Run 1 (Case Study 2): Summary of Uniqueness results. Kappa and ICC are relative to the expert-expert baseline of 0.42 and 0.43 respectively. MAE values compare to the observed Expert-Expert MAE = 1.12. Mean Diff ± 1.96 SD reports the 95% limits of agreement from a Bland-Altman analysis; this reflects the spread of itemwise differences. "Equiv?" indicates whether TOST found AI-expert equivalence within ±1.0 margin. Bold indicates model is within 20% of expert-expert agreement.

	Comparison	Kappa	ICC	MAE	Mean Diff ± 1.96 SD	Equiv?	Spear.	Wilcoxon p-corr	AUC	Tests Passed
	Expert 1 vs Expert 2	0.42	0.43	1.12	−0.81 ± 2.32	-	0.57	$5.53 \cdot 10^{-10}$	0.68	-
Expert 1	Expert 3	**0.36**	**0.36**	**1.09**	**0.79 ± 2.34**	**True**	**0.46**	$9.11 \cdot 10^{-10}$	**0.64**	8/9
	AI Judge: No Context	0.16	0.16	**1.32**	1.21 ± **2.11**	False	0.30	$4.36 \cdot 10^{-16}$	**0.61**	3/9
	AI Judge: Text	0.27	0.27	**1.05**	**0.70 ± 2.34**	**True**	0.36	$3.66 \cdot 10^{-08}$	**0.63**	5/9
	AI Judge: Text + Image	**0.37**	**0.37**	**0.93**	**0.40 ± 2.20**	**True**	0.40	$6.23 \cdot 10^{-04}$	**0.66**	7/9
	AI Judge: T + I + Reasoning	0.26	0.26	**0.98**	**0.24 ± 2.65**	**True**	0.34	$\mathbf{5.44 \cdot 10^{-02}}$	**0.66**	6/9
	AI Judge: o3 Reasoning	0.26	0.26	1.41	1.19 ± **2.46**	False	0.43	$2.83 \cdot 10^{-14}$	**0.66**	2/9
Expert 2	Expert 3	**0.66**	**0.67**	**0.83**	**-0.03 ± 2.15**	**True**	**0.68**	$\mathbf{1.00 \cdot 10^{+00}}$	**0.69**	9/9
	AI Judge: No Context	**0.41**	**0.40**	**1.00**	**0.40 ± 2.43**	**True**	0.45	$1.76 \cdot 10^{-03}$	**0.63**	7/9
	AI Judge: Text	**0.50**	**0.50**	**0.97**	**-0.12 ± 2.46**	**True**	**0.51**	$\mathbf{1.00 \cdot 10^{+00}}$	**0.64**	9/9
	AI Judge: Text + Image	**0.44**	**0.44**	**1.06**	**-0.41 ± 2.54**	**True**	**0.48**	$2.19 \cdot 10^{-03}$	**0.64**	8/9
	AI Judge: T + I + Reasoning	**0.40**	**0.40**	**1.13**	**-0.57 ± 2.74**	**True**	**0.47**	$6.95 \cdot 10^{-05}$	**0.64**	8/9
	AI Judge: o3 Reasoning	**0.55**	**0.55**	**0.95**	**0.38 ± 2.46**	**True**	**0.55**	$6.44 \cdot 10^{-03}$	**0.65**	8/9

other experts and with o3. In contrast, o3 aligns quite closely with Experts 2 and 3.

Furthermore, for technical goodness, the Kruskal-Wallis tests showed mixed evidence of group differences across iterations. In runs 1 and 2, significant omnibus effects were detected ($H = 30.6$, $p = 1.1 \times 10^{-6}$; $H = 16.4$, $p = 9.6 \times 10^{-4}$), whereas run 6 showed no overall group differences ($H = 6.3$, $p = 0.096$).

Post hoc comparisons reveal that o3 Reasoning consistently assigned higher technical goodness ratings than the experts, reflected in negative Cliff's δ values (≈ -0.14 to -0.34). These differences were strongest relative to Expert 1 and Expert 3 (large effect sizes, corrected $p < 0.01$), and more modest or nonsignificant relative to Expert 2. By contrast, the experts did not significantly differ from one another in any run (all corrected $p > 0.20$, δ near 0).

5.2.3 Trends among design metrics. Comparing across the two metrics, AI judges consistently achieved higher agreement with experts for uniqueness than for technical goodness. Nearly all models passed more tests on uniqueness, with top AI judges approaching or even matching expert-expert agreement levels (e.g., Text and o3 Reasoning vs. Expert 2). In contrast, technical goodness proved to be a more challenging metric for AI judges, with the best models typically passing only 5-7 tests on average compared to perfect 9/9 alignment among the experts themselves. This suggests that while AI judges capture subjective evaluations of uniqueness relatively well, the more domain-specific and process-dependent metric of technical goodness remains more difficult to replicate.

Table 7 Uniqueness (Case Study 2): Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Expert 3	8	7	8	7.67/9
	AI Judge: No Context	3	2	3	2.67/9
	AI Judge: Text	5	5	5	5/9
	AI Judge: Text + Img.	7	7	8	7.33/9
	AI Judge: Txt. + Img. + Reas.	6	6	5	5.67/9
	AI Judge: o3 Reasoning	2	3	5	3.33/9
Expert 2	Expert 3	9	9	9	9/9
	AI Judge: No Context	7	3	8	6/9
	AI Judge: Text	9	9	8	8.67/9
	AI Judge: Text + Img.	8	8	8	8/9
	AI Judge: Txt. + Img. + Reas.	8	4	1	4.33/9
	AI Judge: o3 Reasoning	8	8	9	8.33/9

Table 8 Technical Goodness (Case Study 2): Summary of Tests Passed Across Runs. All runs are out of a total of 9 tests.

	Comparison	Run 1	Run 2	Run 3	Avg.
Expert 1	Expert 3	9	9	9	9/9
	AI Judge: No Context	4	4	5	4.33/9
	AI Judge: Text	4	4	4	4/9
	AI Judge: Text + Img.	4	4	4	4/9
	AI Judge: Txt. + Img. + Reas.	2	2	2	2/9
	AI Judge: o3 Reasoning	4	4	7	5/9
Expert 2	Expert 3	9	9	9	9/9
	AI Judge: No Context	4	4	4	4/9
	AI Judge: Text	5	9	5	6.33/9
	AI Judge: Text + Img.	4	7	5	5.33/9
	AI Judge: Txt. + Img. + Reas.	2	3	3	2.67/9
	AI Judge: o3 Reasoning	6	7	9	7.33/9

6 Discussion

Our results demonstrate that reasoning-enabled vision-language models equipped with in-context learning can closely approximate, and at times exceed, expert performance in subjective design evaluation. Across both case studies, trends emerged that can inform when and how AI judges achieve expert-level equivalence, and the implications for future design research.

6.1 Reasoning Models Lead Performance. Reasoning-enabled models, particularly the Text + Image + Reasoning and o3 Reasoning AI judges, consistently ranked as top performers across design metrics. In nearly all trials, one or both of these reasoning judges placed first or second among the five AI judges. This finding highlights the importance of combining multimodal inputs with structured reasoning, suggesting that prompting strategies which encourage deliberative output can substantially improve alignment with expert ratings. In contrast, AI Judge: No Context

showed the weakest alignment with expert-level agreement for uniqueness and creativity, highlighting the value of few-shot demonstrations via ICL.

This trend is particularly evident in Tables 3 and 2, where the four ICL-based judges consistently outperformed the No Context model in aligning with expert ratings for uniqueness and creativity. Interestingly, however, AI Judge: No Context outperformed many of the others for drawing quality. For drawing quality, this may be because the model was still given the target image in the query, suggesting that the VLM's pre-trained understanding of visual quality may suffice without demonstrations. This aligns with findings from [18], which reported that drawing quality was the only design metric where image input improved AI predictions more than text descriptions.

In some cases, AI Judge: Text performed as well as, or even better than, AI Judge: Text + Image, particularly for creativity, usefulness, and drawing quality as well as in Expert 2 comparisons for uniqueness. This result is somewhat counterintuitive, as one might expect that additional contextual information (i.e., the image) would enhance performance. One possible explanation is that the experts may have relied more heavily on the textual descriptions than the sketches when making their assessments. This aligns with findings from [18], which showed that text descriptions had greater predictive power than sketches. Future work could further investigate this phenomenon.

6.2 Comparison with Trained Novices. In Case Study 1, the top AI judges rivaled or surpassed trained novices in multiple design metrics. For uniqueness, creativity, and drawing quality, the best AI judges matched or beat two out of three novices in the number of statistical tests passed. For usefulness, which proved most difficult, the top AI judge surpassed all three novices when compared to one expert, but only one novice when compared to the second expert. These results indicate that reasoning-enabled VLMs can serve as viable substitutes for trained novices, reducing the cost and time burden of human evaluations while maintaining comparable reliability.

6.3 Dependence on the Reference Expert. Performance varied substantially depending on which expert served as the comparison baseline. In Case Study 1, for creativity, and drawing quality, both AI judges and trained novices tended to achieve higher alignment with Expert 2 than with Expert 1. Furthermore, in Case Study 2 the AI Judges tended to achieve higher alignment with Expert 2 than Expert 1, as described in Section 5.2.2. Future research might explore what factors make a rater's judgments more or less "learnable" or replicable by an AI judge. These might include tendency toward a mean value rating, self-consistency, or even years of experience. Alternatively, this may be more impacted by features of the AI judge, such as the selected VLM or the type of context provided. While we do not answer these questions in this work, we do note in Section 5.2.2 that in Case Study 2, the AI judge better matches the ratings of Experts 2 and 3 than Expert 1, which aligns with the differences seen just among the three experts: Post hoc Mann-Whitney U tests show that Expert 1 consistently diverges from both Expert 2 and Expert 3, with moderate effect sizes. So perhaps if an expert diverges from other experts' ratings, we can expect that an AI judge will be less able to learn their preferences. In practice, understanding how an expert rates designs may shed insight on how effective an AI judge will be at learning and replicating those ratings.

6.4 Exceeding Expert-Expert Agreement. In several instances, AI judges achieved stronger agreement with one expert than the experts did with each other. For example, in Case Study 2's uniqueness trials and Case Study 1's drawing quality trials, reasoning-enabled AI judges outperformed the expert-expert baseline in terms of Kappa and ICC while also achieving lower mean absolute error and mean difference. These cases demonstrate that with just a few in-context learning examples, reasoning-enabled VLMs can produce subjective ratings that match an expert's ratings at or beyond expert-level equivalence.

6.5 Metric-Dependent Differences in AI and Expert Ratings. In Case Study 1, the Kruskal–Wallis analysis and post hoc pairwise tests reveal systematic, metric-dependent rating shifts between the top performing AI Judge, o3 Reasoning, and human experts. Specifically, there exist design metrics (i.e. creativity) for which AI judge consistently assigns higher ratings than both experts, as well as metrics for which the AI judge consistently assigns lower ratings than both experts. These trends are summarized in Section 5.1.5.

cross all runs, the o3 Reasoning AI judge rates designs higher than both experts for creativity and usefulness, while assigning lower ratings than both experts for uniqueness. For drawing quality, a different pattern emerges: Expert 1 consistently acts as the high rater, with the *o3 Reasoning* judge assigning lower ratings than Expert 1 and exhibiting no statistically distinguishable difference from Expert 2.

As discussed in Section 2.2, there are various examples in the literature of biases exhibited by AI judges, such as verbosity, position, and style bias [25]. However, here we are looking at a different type of bias: a systematic shift away from expert ratings. Prior research compared AI judges and humans in peer reviewing manuscripts, and found that across 22 manuscripts, AI rejection rates were statistically significantly lower than humans, meaning AI judges were less harsh [66]. Another example of AI raters being less harsh than humans was in Goshtasbi et al.'s 2024 study on facial attractiveness, in which AI-based tools' ratings of facial attractiveness were significantly higher than those of humans in a study with a 40-photograph dataset [67]. However, a different study compared AI raters to humans for judging 30 essays written by English as a foreign language students, and found that AI judges consistently graded lower than human judges, meaning AI judges were more harsh [68]. The impact of subjectivity vs objectivity is also expressed by Ragolane et al. who tested GPT-3.5 and GPT-4 against human raters for grading educational assignments and concluded that automated AI tools were better able to match human ratings when clear instructions and criteria were provided, but struggled in matching humans when tasks were subjective or vague [69]. These conflicting findings, and the small sample size in each study, indicate that there is not yet field-wide consensus on how an AI judge will differ from a human judge.

We hypothesize that for subjective and positive metrics, AI judges may give ratings higher than human raters, as a sort of "flattery bias." Meanwhile for objective, well-defined, or neutral-to-negative metrics we may not see the same trend. Future work should explore where and why we see an AI judge rate things systematically higher or lower than a human.

6.6 Patterns Across Statistical Tests. Examining results across the nine statistical tests reveals systematic patterns. The Two One-Sided Tests (TOST) for equivalence and Jaccard AUC tests were most often passed, suggesting that AI judges reliably capture overall equivalence and top-set selection. By contrast, the Wilcoxon signed-rank test was least often passed, reflecting the challenge of replicating fine-grained distributional characteristics of expert ratings. Weighted Cohen's Kappa and ICC tended to move together: a judge either succeeded on both or failed on both, reinforcing their joint role in measuring reliability.

6.7 Expert-Expert Agreement Does Not Guarantee AI-Expert Agreement. An important finding is that high expert-expert agreement in a metric does not necessarily translate into high AI or novice performance. In Case Study 1, usefulness was the metric with the strongest expert alignment (highest Kappa, ICC, Spearman, and AUC, and lowest error) yet both AI judges and novices performed poorest on this metric. AI judges typically passed only

2-4 tests against Expert 1 and 1-2 tests against Expert 2. This paradox may suggest that experts apply nuanced, domain-specific heuristics to usefulness that are not easily conveyed through a small set of demonstrations, making it difficult for both trained novices and AI to replicate.

6.8 Implications for Design Research. Our results show that *large, general vision-language models* can approach and sometimes exceed expert-level performance on subjective design metrics with minimal overhead, provided that we carefully craft the context. Moreover, reasoning-enabled VLM judges show potential to outperform or match trained novices in design evaluation tasks. This is a powerful finding for design scenarios with limited data or budget to collect massive labels for training. However, performance is metric- and expert-dependent, emphasizing the need for multi-dimensional validation before deployment. The consistent advantage of reasoning models suggests that future AI judges should incorporate reasoning to align more closely with expert ratings. Furthermore, the context has a large impact on performance, and future research should comprehensively explore how to select the best context.

AI judges' strong performance when compared to trained novices has promising cost and time saving implications for design evaluations. These findings also suggest that AI judges could be used in place of trained novices as proxies for expert raters. As a showcase of the potential savings, let us assume an expert is paid \$200/hour to rate designs based on uniqueness. Each design takes the Expert 15 seconds to rate, so 1,000 designs would cost $\$200 \times 4.167$ hours = \$833.33. With the ICL framework, an expert could rate just 9 designs, totaling \$12.50. From those 9 designs, we can use a reasoning model to rate the remainder of the 1,000. In our 12 trials with OpenAI's o1, running the reasoning model cost ~ \$0.17/ design, so for 991 additional designs, it would cost \$168.47. Therefore, the total cost savings are \$833.33 – \$12.50 – \$168.47 = \$652.46, or **78% cost savings**.

This value becomes even clearer when considering the cost and time of hiring experts and the opportunity cost of using their time for repetitive ratings. Unlike human raters, who experience rating fatigue and shifts in judgment [70], AI provides consistent evaluations at scale. Lastly, the cost of LLMs and VLMs is steadily decreasing[71,72], especially with advancements in open-source models[13], making AI-assisted evaluation increasingly viable.

6.9 Implications for Human-AI Design Workflows. Building on the implications described above, this work informs how AI judges might be used in a human-AI design workflow or even an agentic workflow. We can imagine an AI judge agent which automatically rates and then ranks a large set of designs, such that human raters can focus their rating efforts on high-quality sets of designs. This sort of AI ranking or filtering in a human-AI workflow has been shown as successful in prior engineering design research [73,74].

In this ranking setup, absolute rating differences between AI and human judges are not as important as high correlation. Furthermore, if we only care about identifying the top performing designs, then statistical metrics like Jaccard similarity (or other measures of top-set overlap) become most important. It is clear that the nature of the workflow informs which statistical tests are most relevant, and vice-versa, which statistical tests have the highest performance may inform the nature of the workflow. Future work that predicts how an AI judges will rate relative to expert raters for a given metric or task will be critical for identifying the most effective human-AI design workflows.

6.10 Statistical Rigor Matters. We emphasize that it is insufficient to rely on a single metric (e.g., correlation or a p-value from a difference test) to claim human-level performance. Our evaluation integrates multiple statistical approaches to provide a comprehensive assessment of AI judge reliability. Each statistical test captures a different aspect of agreement; some focus on absolute error, while others assess ranking or distributional similarity. Passing one test does not necessarily imply passing others, highlighting the need for a multifaceted evaluation framework.

To exemplify this, for the drawing quality metric, Trained Novice 3 frequently passed the TOST equivalence test and exhibited low MAE when compared to Expert 1, yet failed in weighted Cohen's Kappa, ICC, and Spearman correlation. This suggests that while their ratings were numerically close on average, their ranking patterns and relative consistency differed significantly. Such discrepancies highlight the limitations of single-metric evaluations and motivate our use of a suite of nine statistical tests to capture different facets of expert equivalence. By incorporating multiple measures, we ensure a more rigorous assessment of whether a model or rater truly aligns with expert judgment across both absolute agreement and ranking consistency.

Our approach also emphasized selecting appropriate parameters and thresholds for the design context. For example, ± 1.0 equivalence margin in our TOST analysis and the 20% error margin from expert-expert agreement were determined by the Likert scale used and expert agreement in our dataset. Future work may explore ways to determine the best parameters for a given application. This flexibility allows for tailoring AI evaluation to different levels of precision required in design assessment tasks.

7 Limitations and Future Work

Though encouraging, these findings are only for two case studies using early-stage sketches. These AI judges should be tested on other design modalities (e.g., CAD models, design prototypes) to confirm generality. Because in-context learning can fail if the domain is too far from the model's pretraining, we might not always see such strong results. Another limitation is interpretability: the AI judge's rating might match an expert's but not necessarily provide a rationale. Future efforts can study the "chain-of-thought" text so the model explains how it judged certain design aspects.

We plan to further study whether the trends among design metrics hold for other datasets. Are uniqueness and drawing quality consistently the metrics for which AI judges can best match experts? Is usefulness consistently difficult? This also raises the question of whether these results would hold when using different vision-language models, particularly open-source alternatives. Our future work will explore this in depth.

A further consideration is the reliability of AI-based evaluation methods when *experts themselves* diverge on certain design aspects. Since experts may disagree on criteria such as creativity or functionality of a given design, when and how should an AI model predict expert ratings? Future work should examine inter-expert variability, perhaps using ensemble AI methods or probabilistic models to capture the distribution of expert ratings and provide uncertainty estimates in AI judgments.

There are important ethical considerations to be made around the use of AI as an evaluator. We aim to highlight the importance of ensuring that an AI's evaluations align with human judgment to mitigate risk, but other factors must also be considered. For instance, an AI judge could amplify expert biases at scale. Moreover, while an AI model may outperform a novice in replicating expert ratings, this does not diminish the value of novice training. It remains crucial to determine which skills engineers and designers should retain and develop, regardless of AI's capabilities. Additionally, we must examine when and how large-scale design evaluation should occur. While this has been explored for decades, such as in Google's Project 10 to the 100th, AI has dramatically increased the volume of generated designs, raising new questions about its role in evaluating this growing set.

8 Conclusion

We proposed using ICL-enhanced vision-language models, which do not undergo additional training, to rate design sketches on subjective metrics like uniqueness, usefulness, and creativity.

To thoroughly verify when such an AI truly performs at a "human expert" level, we combined multiple statistical tests: inter-rater agreement, error analysis, distributional checks, correlation, top-set overlap, *and* equivalence (TOST). This ensures that claims of "no difference" are supported by *actual evidence of equivalence* rather than the mere absence of significance. Our findings suggest that AI judges can serve as viable substitutes for trained novices, outperforming them in several cases and providing scalable, cost-efficient alternatives for large-scale evaluations. More importantly, with carefully constructed few-shot prompting, reasoning-supported VLMs achieved expert-equivalent performance for design metrics such as uniqueness and drawing quality. Furthermore, we find that for specific statistical measures like mean absolute error and Spearman's rank correlation coefficient, reasoning-enabled AI judges can exceed expert-expert levels of equivalence when rating creativity or drawing quality. We conclude that *both* in-context learning and rigorous multi-metric, multi-test statistical validation are crucial to confidently deploy AI judges in design. Moving forward, we believe this holistic approach will generalize to other subjective domains, guiding practitioners to adopt AI safely and effectively as a peer to human experts.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2231254 and the NSF Graduate Research Fellowship.

References

[1] Christiaans, H., 2002, "Creativity as a Design Criterion," Creativity Research Journal - CREATIVITY RES J, **14**, pp. 41–54.
[2] Christensen, B. T. and Ball, L. J., 2016, "Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments," Design Studies, **45**, pp. 116–136.
[3] Sternberg, R. J., 1999, *Handbook of creativity*, Cambridge University Press.
[4] Baer, J., Kaufman, J. C., and Gentile, C. A., 2004, "Extension of the consensual assessment technique to nonparallel creative products," Creativity research journal, **16**(1), pp. 113–117.
[5] Baer, J., 2015, "The importance of domain-specific expertise in creativity," Roeper Review, **37**(3), pp. 165–178.
[6] Kaufman, J. C., Plucker, J. A., and Baer, J., 2008, *Essentials of creativity assessment*, Vol. 53, John Wiley & Sons.
[7] Kaufman, J. C., Baer, J., and Cole, J. C., 2011, "Expertise, domains, and the consensual assessment technique," The Journal of creative behavior, **43**(4), pp. 223–233.
[8] Ambile, T. M., 1996, *Creativity in Context*, Westview Press, Boulder, Colorado.
[9] Miller, S. R., Hunter, S. T., Starkey, E., Ramachandran, S., Ahmed, F., and Fuge, M., 2021, "How Should We Measure Creativity in Engineering Design? A Comparison Between Social Science and Engineering Approaches," Journal of Mechanical Design, **143**(3), p. 031404.
[10] Lykourentzou, I., Ahmed, F., Papastathis, C., Sadien, I., and Papangelis, K., 2018, "When Crowds Give You Lemons: Filtering Innovative Ideas using a Diverse-Bag-of-Lemons Strategy," Proc. ACM Hum.-Comput. Interact., **2**(CSCW).
[11] 2021, *Predicting a Paradigm Shift: Exploring the Relationship Between Cognitive Style and the Paradigm-Relatedness of Design Solutions*, Vol. Volume 6: 33rd International Conference on Design Theory and Methodology (DTM) of International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.
[12] Burnap, A., Ren, Y., Gerth, R., Papazoglou, G., Gonzalez, R., and Papalambros, P. Y., 2015, "When crowdsourcing fails: A study of expertise on crowdsourced design evaluation," Journal of Mechanical Design, **137**(3), p. 031101.
[13] Liu, H., Li, C., Wu, Q., and Lee, Y. J., 2023, "Visual Instruction Tuning," 2304.08485, https://arxiv.org/abs/2304.08485
[14] OpenAI, 2024, "OpenAI o1 System Card," Accessed: 2025-03-17, https://cdn.openai.com/o1-system-card.pdf
[15] OpenAI, 2024, "GPT-4o System Card," arXiv preprint arXiv:2410.21276, Accessed: 2025-03-17.
[16] Zhao, R., Yuan, Q., Li, J., Fan, Y., Li, Y., and Gao, F., 2024, "DriveLLaVA: Human-Level Behavior Decisions via Vision Language Model," Sensors, **24**(13).
[17] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., and Ramanan, D., 2024, "Evaluating Text-to-Visual Generation with Image-to-Text Generation," 2404.01291, https://arxiv.org/abs/2404.01291
[18] Song, B., Miller, S., and Ahmed, F., 2023, "Attention-Enhanced Multimodal Learning for Conceptual Design Evaluations," Journal of Mechanical Design, **145**(4), p. 041410.
[19] Edwards, K. M., Peng, A., Miller, S. R., and Ahmed, F., 2021, "If a Picture is Worth 1000 Words, Is a Word Worth 1000 Features for Design Metric Estimation?" Journal of Mechanical Design, **144**(4), p. 041402.
[20] Su, H., Song, B., and Ahmed, F., 2023, "Multi-modal Machine Learning for Vehicle Rating Predictions Using Image, Text, and Parametric Data," *Proceedings of the International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, ASME, Boston, MA.
[21] Yuan, C., Marion, T., and Moghaddam, M., 2021, "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model," Journal of Mechanical Design, **144**(2), p. 021403.
[22] Picard, C., Edwards, K. M., Doris, A. C., Man, B., Giannone, G., Alam, M. F., and Ahmed, F., 2024, "From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design," 2311.12668, https://arxiv.org/abs/2311.12668
[23] Ni, B. and Buehler, M. J., 2024, "MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge," Extreme Mechanics Letters, **67**, p. 102131.
[24] Jadhav, Y. and Farimani, A. B., 2024, "Large Language Model Agent as a Mechanical Designer," 2404.17525, https://arxiv.org/abs/2404.17525
[25] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I., 2023, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Neural Information Processing Systems (NeurIPS) 2023, Datasets and Benchmarks Track*, NeurIPS.
[26] Alipour, L., Faizi, M., Moradi, A. M., and Akrami, G., 2017, "The impact of designers' goals on design-by-analogy," Design Studies, **51**, pp. 1–24.
[27] Cheng, P., Mugge, R., and Schoormans, J. P., 2014, "A new strategy to reduce design fixation: Presenting partial photographs to designers," Design Studies, **35**(4), pp. 374–391.
[28] Chan, J., Dow, S. P., and Schunn, C. D., 2018, "Do the best design ideas (really) come from conceptually distant sources of inspiration?" *Engineering a Better Future*, Springer, Cham, pp. 111–139.
[29] Galati, F., 2015, "Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity," Creativity Research Journal, **27**(1), pp. 24–30.
[30] Amabile, T. M., 1982, "Social psychology of creativity: A consensual assessment technique." Journal of personality and social psychology, **43**(5), p. 997.
[31] Baer, J. and Kaufman, J. C., 2019, "Assessing creativity with the consensual assessment technique," *The Palgrave Handbook of Social Creativity Research*, Springer, pp. 27–37.
[32] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for measuring ideation effectiveness," Design studies, **24**(2), pp. 111–134.
[33] Linsey, J. S., Green, M. G., Murphy, J. T., Wood, K. L., and Markman, A. B., 2005, ""Collaborating To Success": An Experimental Study of Group Idea Generation Techniques," *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 4742, pp. 277–290.
[34] John, B. and Sharon, S. M., 2009, *Assessing Creativity Using the Consensual Assessment Technique*, IGI Global, Hershey, PA, USA, pp. 65–77.
[35] Cseh, G. M. and Jeffries, K. K., 2019, "A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research." Psychology of Aesthetics, Creativity, and the Arts, **13**(2), p. 159.
[36] LeBreton, J. M. and Senter, J. L., 2008, "Answers to 20 Questions About Interrater Reliability and Interrater Agreement," Organizational Research Methods, **11**(4), pp. 815–852.
[37] James, L. R., Demaree, R. G., and Wolf, G., 1984, "Estimating Within-Group Interrater Reliability With and Without Response Bias," Journal of Applied Psychology, **69**(1), p. 85.
[38] Lance, C. E., Butts, M. M., and Michels, L. C., 2006, "The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?" Organizational Research Methods, **9**(2), pp. 202–220.
[39] ASME, 2021, *Design Form and Function Prediction From a Single Image*, Vol. Volume 2: 41st Computers and Information in Engineering Conference (CIE) of International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.
[40] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., 2020, "Language models are few-shot learners," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA.
[41] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J., 2025, "A Survey on LLM-as-a-Judge," 2411.15594, https://arxiv.org/abs/2411.15594
[42] Hu, R., Cheng, Y., Meng, L., Xia, J., Zong, Y., Shi, X., and Lin, W., 2025, "Training an LLM-as-a-Judge Model: Pipeline, Insights, and Practical Lessons," *Companion Proceedings of the ACM on Web Conference 2025*, ACM, p. 228–237, doi: 10.1145/3701716.3715265.
[43] Chiang, C.-H., Lee, H.-y., and Lukasik, M., 2025, "TRACT: Regression-Aware Fine-tuning Meets Chain-of-Thought Reasoning for LLM-as-a-Judge," *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds., Association for Computational Linguistics, Vienna, Austria, pp. 2934–2952, doi: 10.18653/v1/2025.acl-long.147.
[44] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I., 2024, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," 2403.04132, https://arxiv.org/abs/2403.04132

[45] Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., Lin, Z., Su, W., Ge, T., Zheng, B., and Ouyang, W., 2024, "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, p. 7421–7454, doi: 10.18653/v1/2024.acl-long.401.

[46] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2018, "Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel," Journal of Mechanical Design, **141**(2), p. 021102.

[47] Luo, M., Xu, X., Liu, Y., Pasupat, P., and Kazemi, M., 2024, "In-context Learning with Retrieved Demonstrations for Language Models: A Survey," 2401.11624, https://arxiv.org/abs/2401.11624

[48] Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z., 2024, "A Survey on In-context Learning," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., Association for Computational Linguistics, Miami, Florida, USA, pp. 1107–1128, doi: 10.18653/v1/2024.emnlp-main.64.

[49] Cohen, J., 1960, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, **20**(1), pp. 37–46.

[50] Fleiss, J. L. and Cohen, J., 1973, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," Educational and Psychological Measurement, **33**(3), pp. 613–619.

[51] Shrout, P. E. and Fleiss, J. L., 1979, "Intraclass correlations: Uses in assessing rater reliability," Psychological Bulletin, **86**(2), pp. 420–428.

[52] Chai, T. and Draxler, R. R., 2014, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," Geoscientific Model Development, **7**(3), pp. 1247–1250.

[53] Bland, J. M. and Altman, D. G., 1986, "Statistical methods for assessing agreement between two methods of clinical measurement," The Lancet, **327**(8476), pp. 307–310.

[54] Conover, W. J., 1999, *Practical Nonparametric Statistics*, 3rd ed., Wiley.

[55] Wilcoxon, F., 1945, "Individual comparisons by ranking methods," Biometrics Bulletin, **1**(6), pp. 80–83.

[56] Spearman, C., 1904, "The proof and measurement of association between two things," The American Journal of Psychology, **15**(1), pp. 72–101.

[57] Schuirmann, D. J., 1987, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," Journal of Pharmacokinetics and Biopharmaceutics, **15**(6), pp. 657–680.

[58] Jaccard, P., 1901, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," Bulletin de la Société Vaudoise des Sciences Naturelles, **37**, pp. 547–579.

[59] Li, X., Sun, Y., and Sha, Z., 2024, "LLM4CAD: Multimodal Large Language Models for Three-Dimensional Computer-Aided Design Generation," Journal of Computing and Information Science in Engineering, **25**(2), p. 021005.

[60] Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S., 2024, "SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models," 2406.01584, https://arxiv.org/abs/2406.01584

[61] Starkey, E. M., Hunter, S. T., and Miller, S. R., 2019, "Are creativity and self-efficacy at odds? an exploration in variations of product dissection in engineering education," Journal of Mechanical Design, **141**(1).

[62] Toh, C. A. and Miller, S. R., 2016, "Choosing creativity: the role of individual risk and ambiguity aversion on creative concept selection in engineering design," Research in Engineering Design, **27**, pp. 195–219.

[63] Starkey, E., Toh, C. A., and Miller, S. R., 2016, "Abandoning creativity: The evolution of creative ideas in engineering design course projects," Design Studies, **47**, pp. 47–72.

[64] Prabhu, R., Miller, S. R., Simpson, T. W., and Meisel, N. A., 2019, "Complex Solutions for Complex Problems? Exploring the Role of Design Task Choice on Learning, Design for Additive Manufacturing Use, and Creativity," Journal of Mechanical Design, **142**(3), p. 031121.

[65] Redmond, M. R., Mumford, M. D., and Teach, R., 1993, "Putting creativity to work: Effects of leader behavior on subordinate creativity," Organizational behavior and human decision processes, **55**(1), pp. 120–151.

[66] Joachim, M. V., Dodson, T. B., and Laviv, A., 2025, "How Artificial Intelligence Differs From Humans in Peer Review," Journal of Oral and Maxillofacial Surgery, **83**(8), pp. 1040–1050.

[67] Goshtasbi, K., Hakimi, A. A., and Wong, B. J., 2024, "Artificial Intelligence Versus Human Focus Group Rating of Facial Attractiveness," Facial Plastic Surgery & Aesthetic Medicine, **26**(4), pp. 371–376, PMID: 38377584.

[68] Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., and Almegren, R. M., 2025, "Evaluating the quality of AI feedback: A comparative study of AI and human essay grading," Innovations in Education and Teaching International, **62**(6), pp. 1858–1873.

[69] Ragolane, M., Patel, S., and Salikram, P., 2024, "AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education," Asian Journal of Education and Social Studies, **50**(10), p. 244–263.

[70] Ling, G., Mollaun, P., and Xi, X., 2014, "A study on the impact of fatigue on human raters when scoring speaking responses," Language Testing, **31**, pp. 479–499.

[71] Korinek, A., 2023, "Generative AI for Economic Research: Use Cases and Implications for Economists," Journal of Economic Literature, **61**(4), p. 1281–1317.

[72] Cottier, B., Snodin, B., and Owen, D., 2025, "LLM inference prices have fallen rapidly but unequally across tasks," Accessed: 2025-03-17, https://epoch.ai/data-insights/llm-inference-price-trends

[73] Edwards, K. M., 2022, "Accelerating the Design Process Through Natural Language Processing-based Idea Filtering," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

[74] Edwards, K. M., Song, B., Porciello, J., Engelbert, M., Huang, C., and Ahmed, F., 2024, "ADVISE: Accelerating the Creation of Evidence Syntheses for Global Development Using Natural Language Processing-Supported Human-Artificial Intelligence Collaboration," Journal of Mechanical Design, **146**(5), p. 051404.