Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis

ANONYMOUS AUTHOR(S)

Generative artificial intelligence (GenAI) is increasingly used to support a wide range of human tasks, yet empirical evidence on its effect on creativity remains scattered. Can GenAI generate ideas that are creative? To what extent can it support humans in generating ideas that are both creative and diverse? In this study, we conduct a meta-analysis to evaluate the effect of GenAI on the performance in creative tasks. For this, we first perform a systematic literature search, based on which we identify $N = 28$ relevant studies for inclusion in our meta-analysis. We then compute standardized effect sizes based on Hedges' g. We compare different outcomes: (i) how creative GenAI is; (ii) how creative humans augmented by GenAI are; and (iii) the diversity of ideas by humans augmented by GenAI. Our results show no significant difference in creative performance between GenAI and humans ($g = -0.05$), while humans collaborating with GenAI significantly outperform those working without assistance ($g = 0.27$). However, GenAI has a significant negative effect on the diversity of ideas for such collaborations between humans and GenAI ($g = -0.86$). We further analyze heterogeneity across different GenAI models (e.g., GPT-3.5, GPT-4), different tasks (e.g., creative writing, ideation, divergent thinking), and different participant populations (e.g., laypeople, business, academia). Overall, our results position GenAI as an augmentative tool that can support, rather than replace, human creativity—particularly in tasks benefiting from ideation support.

+++ Attention +++

- -

Author 1 had his Overleaf editor session open before the Overleaf outage on the day of conference deadline and continued working during the outage without realizing there was a problem. As a result, important changes to the library and discussion of RQ 2b made in this session were overwritten. This was not realized by the authors until after the extended deadline of 12:00 PM EST. We have incorporated these changes into this version. We apologize for any inconvenience this may have caused.

---

1 INTRODUCTION

Generative artificial intelligence (GenAI) refers to a class of machine learning technologies that have the capability to generate new content that resembles human-created output, such as images, text, audio, and videos [21]. GenAI can thus support various human tasks such as writing, software development, composing lyrics, and academic research, often at a performance similar to that of humans [7, 12, 26, 28, 50, 72]. On top of that, Generative AI has also become a valuable tool in creative industries—spanning graphic design, advertising, fashion, writing, and visual arts [36, 59].
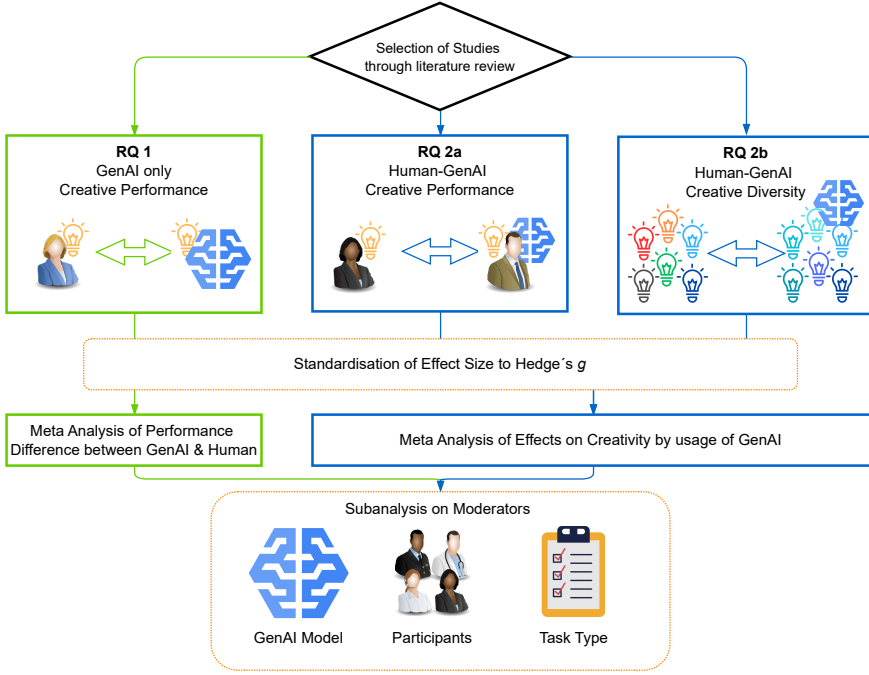

Fig. 1. **Overview of the meta-analysis procedure.** We address three research questions: (RQ1) creative performance of GenAI alone, (RQ2a) creative performance of humans supported by GenAI, and (RQ2b) diversity of ideas in human-GenAI collaboration. After selecting studies via systematic literature review, we standardize effect sizes (Cohen's *d*) and conduct meta-analyses, including a moderator analysis to study heterogeneity across types of GenAI models, tasks, measurements, and participant populations.

Yet, empirical evidence on the benefits of GenAI for creative performance is scattered, and the theoretical arguments are often inconsistent or even contradictory. On the one side, cognitive research argues that creativity is an inherently human trait [2, 51]. One common issue in practice is that GenAI models often lack the tacit knowledge required for systematic, compositional reasoning such as multi-step problem solving to generate ideas perceived as creative, especially in real-world tasks from businesses. Similarly, GenAI is likely to reproduce ideas seen during training rather than ideating novel ideas [30]. Hence, simply by means of the training data of GenAI, the generated ideas may also be less diverse than those of humans [17]. On the other side, there is early evidence suggesting that outputs from GenAI are perceived as being creative [65]. For example, several studies find benefits from GenAI in creative tasks, but these findings are typically limited to specific creative tasks (e.g., story writing, ideating business models) [17, 39, 64, 72]. This may suggest that GenAI can generally improve creativity but it is often unclear how well the results generalize across domains.

Here, we perform a meta-analysis to evaluate the effect of GenAI on creative performance.[1] For this, we focus on different outcomes, namely: (i) how creative GenAI is; (ii) how creative humans

---

augmented by GenAI are; and (iii) the diversity of ideas by humans augmented by GenAI.[2] To this end, we analyze the following **research questions (RQs)**:

- **RQ1:** *How creative are ideas generated by GenAI (compared to humans without GenAI support)?*
- **RQ2a:** *How creative are ideas generated by humans when supported by GenAI (compared to humans without GenAI support)?*
- **RQ2b:** *How diverse are ideas generated by humans when supported by GenAI (compared to humans without GenAI support)?*

To answer the above research questions, we first perform a systematic literature search and then conduct a meta-analysis. Overall, we retrieved $N = 693$ studies for inclusion, and eventually identified $N = 28$ studies with empirical results. We then calculated the standardized effect size Hedges' g and computed a random-effects meta-analysis. We further ran a moderator analysis on heterogeneity across different GenAI models (e.g., GPT-3.5, GPT-4), different tasks (e.g., creative writing, ideation, divergent thinking), and different participant populations (e.g., laypeople, business, and academia) to provide a general but differentiated understanding of the effect of GenAI on creativity.

2 METHODS

In this section, we describe our data collection to identify relevant studies and statistical analysis.

2.1 Search Strategy

Search databases: To ensure transparency, reproducibility, and quality of our meta-analysis, we follow the PRISMA 2020 framework [45] for systematic literature reviews. To identify relevant studies, we searched the following databases: (i) Web of Science, (ii) SSRN, and (iii) arXiv. Our search includes non-peer-reviewed studies to reflect the rapidly evolving nature of GenAI research and to capture recent advances. Our search was limited to publications in the English language from the past five years, which is loosely aligned with the emergence of foundational models such as GPT and BERT. The knowledge cutoff of our search is May 2, 2025.

Search query: Our search query is intentionally broad to include various ways to relate to GenAI technology and creativity. Overall, our search query is inspired by Schemmer et al. (2022) [52], but which we adapted to our research question, namely, creativity:

Search query

```
TITLE("creativity" OR "creative" OR "ideation" OR "idea")
AND
("AI" OR "Artificial Intelligence" OR "LLM" OR "Large Language Models")
```

Inclusion/exclusion: The process for inclusion/exclusion in our systematic literature review is shown in Figure 2 (based on the format of the PRISMA 2020 Flowchart [46]). Literature screening, eligibility checks, and final inclusion were performed by one person (the first author). In the identification phase, $N = 691$ publications were identified by our search query, out of which $n = 96$ duplicate publications were removed manually. Out of the remaining $n = 595$ publications, both

---

[2]We also considered a fourth outcome, namely, the diversity of ideas generated by humans with GenAI support. Yet, our literature search did not return a sufficient number of empirical studies for this outcome. Hence, we refrain from analyzing this outcome. However, we identify this gap as a promising opportunity for future research, which we elaborate on in the discussion section.

the title and abstract were screened for inclusion in the assessment of eligibility. Here, n = 516 publications were excluded due to a lack of fit (e.g., a focus on legal studies).

Subsequently, n = 79 records were assessed for eligibility. Studies were eligible if they:

(1) The study design was aimed at comparing (a) the creativity performance of humans versus GenAI or (b) the creative performance of humans with vs. without GenAI support. The study further followed a between-subject experiment design, which is crucial to make rigorous statistical comparisons.

(2) The study had to report sufficient statistics (e.g., the group means and standard deviations or equivalent statistics) that allow for computation of standardized effect size Hedges' g. In cases where such statistics were not directly reported, we examined whether the publication was accompanied by raw data, either in supplementary materials or associated repositories, that would enable us to reconstruct the necessary statistics. However, if these data were not accompanied by clearly documented analysis scripts and/or the reconstruction would have required substantial interpretation or rewriting of the original code, the study was excluded based on insufficient raw data.

(3) The study had to measure creative performance using an established measurement dimension (e.g., novelty [60], originality [40], diversity [47, 56]).

(4) Do not provide raw data to prevent mistakes when handling unknown data

In this stage, publications were excluded due to (i) insufficient study design (n = 33) (ii) insufficient statistics (n = 8), (iii) insufficient creativity measurement (n = 4), and (iv) insufficient raw data (n = 2). Cases with ambiguity were resolved through discussion until consensus was reached (based on Author #1 together with Author #2 and Author #3).

We contacted the corresponding authors from 15 publications via e-mail to address two of the above reasons for exclusion, namely, (ii) insufficient statistics and (iv) insufficient data reporting. If no response was received after the initial contact, a follow-up e-mail was sent to maximize the inclusion of eligible studies. As a result, we obtained sufficient additional information for inclusion from four studies (n = 4). All studies that did not elicit a response were excluded.

Given the novel scope of the research question and the evolving terminology in the field, we recognized the possibility that the search strategy might overlook relevant studies. To address this, the authors manually included two additional studies that met all eligibility criteria but were not captured by the original database query.

In total, we identified n = 28 publications that met the eligibility criteria for inclusion in our meta-analysis. Several of these publications reported multiple experiments or included multiple outcome measures related to creativity and diversity. As a result, the 28 studies correspond to a total of 127 effect size estimates.

2.2 Data Collection

The n = 28 studies that met the inclusion criteria were recorded in our database. To ensure transparency and reproducibility, our database capturing all relevant study features and outcomes is deposited in our project repository.

An overview of the extracted dimensions is summarized in Table 1, which serves as the basis for exploring heterogeneity in creative performance across various study and task characteristics. Specifically, we extracted study-level data of metadata (title, author, abstract, publication date, source). We further provide details about the experimental design (task type, GenAI type, GenAI model, creativity measurement, measurement evaluator), participant characteristics (recruitment source, professional domain), and outcome statistics (means, SD, SE, n_{total}, n_{control}, $n_{\text{treatment}}$, F-value, standardized β, and $SE - \beta$). We discuss a few relevant dimensions in the following.
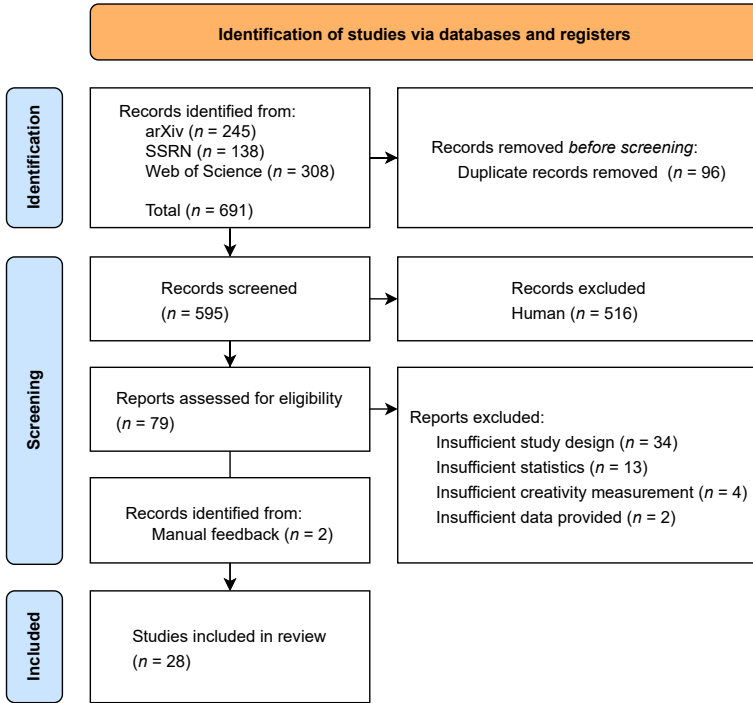
Fig. 2. **PRISMA flow chart.**

The **task type** reflects the notion that different creative tasks tap into distinct aspects of creativity (and thus different cognitive processes), such as compositional skills, imaginative reasoning, or divergent thinking (i.e., the ability to generate multiple, varied ideas in response to an open-ended prompt). The studies in our analysis employ a variety of creative tasks, each targeting different dimensions of creative cognition. For example, the so-called *alternate uses task* (AUT) [27] asks participants to generate unconventional uses for everyday objects (e.g., "a brick"), which assesses originality. The so-called *consequences task (CT)* [67] involves imagining outcomes of improbable scenarios (e.g., "What would happen if people could fly?"), which taps into imaginative thinking and hypothetical reasoning. The so-called *divergent association task (DAT)* [33] requires listing unrelated words (e.g., "apple", "justice", "galaxy"). *Forward flow* (FF) measures [25] the conceptual distance between sequential thoughts in open-ended responses, reflecting the natural flow of creative ideation. Finally, other studies rely upon creative writing tasks such as short story composition or poetry to challenge narrative creativity, but also business ideation tasks (e.g., coming up with a new business model).

Prior work has suggested that creative performance comprises multiple dimensions [40], and **creativity measurements** can be obtained in subjective (e.g., human ratings) or objective (e.g., via metrics from natural language processing such as semantic distance or cosine similarity [22] ways. Subjective assessments commonly rely on scales such as perceived creativity, originality, or novelty—terms that show some overlap conceptually but are occasionally operationalized as distinct constructs [35]. For example, *creativity* is often perceived as a combination of novelty and effectiveness, *originality* refers to how uncommon or unique an idea is, and *novelty* emphasizes newness or unfamiliarity. These dimensions may be interpreted differently by human raters,

depending on context and individual experience, because of which we also coded the identity of the **evaluator**. For example, self-assessments may introduce bias due to over- or underestimation of one's performance, while expert ratings tend to be more consistent and informed. Rule-based or AI-based evaluations offer standardization and scalability but may miss nuanced judgment or (tacit) domain knowledge.

In cases where studies reported multiple measurements per experiment, we prioritized evaluations of creative performance by experts (over self-assessments). When multiple experiments were reported within a single study, each experiment was included separately, with one outcome measurement per experiment. Creative performance was encoded using the most conceptually appropriate metric available—ideally an overall creativity score (if available) or, alternatively, measures of originality or novelty before using any other measurement.[3]

Table 1. **Extracted dimensions.**

Dimensions	Values
GenAI model	GPT-4o, GPT-4, GPT-4all, GPT-3.5-turbo, GPT-3.5, GPT-3, Claude, SparkDesk, Qwen, Dou Bao, alpaca, bing, dolly, koala, oa, stablelm, vicuna, not disclosed (n.d.); *optional:* multiple models
Participants	Academia, business, laypeople, not disclosed
Task type	alternate uses task (AUT), consequences task (CT), divergent associations task (DAT), forward flow (FF), creative writing, creative problem solving, creative thinking, divergent thinking, ideation product, ideation item usage, ideation research proposal, ideation business concepts, ideation other
GenAI type	text-to-text (T2T), text-to-image (T2I)
Recruitment source	University, Prolific, Mturk, Public, Company
Creativity measurement	creativity scale, originality scale, novelty scale, semantic distance, cosine similarity, creative problem-solving scale, flexibility score
Evaluator	self-assessed, laypeople, expert, rule-based, AI

2.3 Statistical Analysis

Hedges' g: For each comparison, we extracted reported Cohen's d or calculated it based on reported statistics [11]. When effect sizes were not directly reported, we converted other statistical measures (e.g., t-values, F-values, means and standard deviations) to Cohen's d using widely accepted conversion formulas [4, 18, 48, 49], which are documented in our project repository. To adjust for potential upward bias in small samples, we applied Hedges' g correction [38]. We report one effect size per individual experiment across all included studies to ensure statistical independence.

Random-effects model: As we expect large heterogeneity regarding treatment, task, and measurement across studies, we chose to estimate a random-effects model as recommended by Cochrane [14]. Pooled effect sizes were estimated under the random-effects model using the DerSimonian-Laird estimator [15] to account for the expected between-study heterogeneity. We calculated 95% prediction intervals to indicate the expected range of true effects across settings.

---

[3]To examine potential heterogeneity across different conceptualizations of creative performance, we conducted exploratory subgroup analyses based on the specific constructs used—namely, creativity, originality, and novelty. However, these analyses revealed no substantial heterogeneity. This may be due to the high conceptual overlap among the constructs of creativity, originality, and novelty, which are often used interchangeably in both academic and applied contexts. We thus omitted the analysis for space but the reproducibility code is available for interested readers in our repository.

Variability (I^2 statistic): The variability in effect sizes across studies was quantified using the I^2 statistic, Cochran's Q test, and by estimating the between-study variance (τ^2) via Jackson's method [32].

Heterogeneity analysis: To further explore potential sources of heterogeneity, we conducted subgroup analyses, such as comparing creativity across participants with different backgrounds (e.g., academic vs. business), to examine differential performance in GenAI-assisted creative tasks. As a robustness check, we also performed meta-regression analyses [62], which yielded results consistent with the subgroup analyses. For reasons of space, detailed meta-regression results are omitted.

Bias assessment: Risk of bias was assessed using the Cochrane Risk-of-Bias 2.0 tool [57], with judgments recorded for selection, performance, detection, and reporting bias. Publication bias was evaluated through Egger's regression test [20] and the trim-and-fill procedure [19] for comparisons including at least ten studies. These analyses revealed no substantial concerns regarding bias; we thus omitted full results for brevity.

Implementation details: All analyses were implemented in R (version 4.2.3) using the `metafor` package (version 4.8-0). Influence diagnostics were conducted via the `influence()` function of the `metafor` package, and leave-one-out sensitivity analyses [63] were used to assess the robustness of pooled estimates. Codes are available in our repository for reproducibility.

8

Overall ID	CP ID	CD ID	HAI ID	Title	GenAI model	Participants	Task type	GenAI type	Recruitment source	Creativity measurement	Evaluator
1			HAI01 R12	A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing [23]	GPT-4, GPT-3.5, bing, claude12, koala, vicuna, oa, bard, GPT-4all, stablelm, dolly, alpaca	Academia	creative writing	T2T	University	creativity scale	Expert
2			HAI02 R6	AI Delivers Creative Output but Struggles with Thinking Processes [70]	GPT-3.5, GPT-4, GPT-4o	Laypeople	AUT	T2T	Public	novelty scale	Human
3	CP01 R6		HAI03 R1	An empirical investigation of the impact of ChatGPT on creativity [39]	GPT-3.5	Laypeople	ideation item usage, ideation product, ideation other	T2T	Mturk/Prolific	creativity scale	Expert
4			HAI04 R1	Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli [24]	GPT-4	Not disclosed	creative interpretation	T2T	Prolific	creativity scale	Human
5			HAI05 R1	Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity [34]	GPT-3, Copy.Ai, Alpa.ai, Studio, YouChat	Academia	AUT	T2T	Prolific	creativity scale	Human
6			HAI06 R1	Best humans still outperform artificial intelligence in a creative divergent thinking task [37]	GPT-3.5, GPT-4, Copy.Ai	Not disclosed	AUT	T2T	Prolific	creativity scale, semantic distance	Human
7			HAI07 R1	Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers [54]	claude-3.5-sonnet	Academia	ideation research ideas	T2T	Public	novelty scale	Expert
8	CP02 R1		HAI08 R2	Creative and Strategic Capabilities of Generative Ai: Evidence from Large-Scale Experiments [43]	GPT-4, Bard	Not disclosed	creative writing	T2T	Prolific	creativity scale	Human
9			HAI09 R4	Creativity and AI [9]	GPT-3.5, GPT-4o	Not disclosed	creative writing, ideation product	T2T	Prolific	creativity scale	Human
10	CP03 R3		HAI10 R1	Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence [41]	n.d.	Academia	creative writing	T2T	Prolific	creativity scale	Expert
11		CD03 R1	HAI11 R1	Evaluating Creative Short Story Generation in Humans and Large Language Models [31]	60 models	Academia	creative writing	T2T	Prolific	creativity scale	Expert
12	CP04 R1			Generative AI Adoption in Human Creative Tasks: Experimental Evidence [73]	GPT-3.5	Academia	ideation other	T2T	University	creativity scale	Human
13	CP05 R2	CD01 R2		Generative artificial intelligence enhances creativity but reduces the diversity of novel content [17]	GPT-4	Not disclosed	creative writing	T2T	Prolific	novelty scale, cosine similarity	Expert / Rule-based
14			HAI12 R1	How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment [3]	GPT 3.5	Not disclosed	AUT	T2T	Public	originality score	Rule-based
15			HAI13 R1	How AI Outperforms Humans at Creative Idea Generation [8]	GPT-4	Business	ideations of new products	T2T	Prolific	creativity scale	Human
16	CP06 R1	CD02 R1		How Experience Moderates the Impact of Generative AI Ideas on the Research Process [16]	GPT-4	Academia	ideation research proposal	T2T	University	novelty scale, cosine similarity	Self-assessed
17	CP07 R1			If ChatGPT can do it, where is my creativity? generative AI boosts performance but diminishes experience in creative writing [42]	GPT-4o, GPT-4-turbo	Academia	creative writing	T2T	Prolific	flexibility score	Human
18	CP08 R1			Interactions with generative AI chatbots: unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving [55]	Dou Bao	Academia	creative problem solving	T2T	University	CPS scale	Rule-based
19	CP09 R1	CD04 R2	HAI14 R1	Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise [10]	GPT-4	Business	ideation business concepts	T2T	Prolific	creativity scale	Human
20	CP10 R1			Large Language Model in Ideation for Product Innovation: An Exploratory Comparative Study [71]	GPT-3.5-turbo	Laypeople	ideation product	T2T	Mturk	novelty scale	Human
21			HAI15 R52	Large Language Models show both individual and collective creativity comparable to humans [58]	GPT-3.5, GPT-4, Claude, Qwen, SparkDesk	Academia	Divergent Thinking, problem solving, creative writing	T2T	University	novelty scale, originality scale	Human
22			HAI16 R1	One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor, 1082–1092 [68]	GPT-4o	Academia	ideation meme content	T2T	Prolific	creativity scale	Human
23	CP11 R1		HAI17 R1	Revolution or inflated expectations? Exploring the impact of generative AI on ideation in a practical sustainability context [1]	GPT-4	Business	ideation other	T2T	Company	novelty scale	Expert
24	CP12 R2			The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing [6]	GPT-4	Business	ideation business concepts	T2T	Public	novelty scale	Expert
25			HAI18 R5	The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks [29]	GPT-4	Academia	AUT / CT / DAT	T2T	Prolific	semantic distance	AI
26			HAI20 R2	The Language of Creativity: Evidence from Humans and Large Language Models [44]	GPT-3	Not disclosed	creative writing	T2T	Prolific	creativity scale	Human
27			HAI21 R3	We're Different, We're the Same: Creative Homogeneity Across LLMs [66]	22 models	Not disclosed	AUT / FF / DAT	T2T	Prolific	semantic distance	Rule-based
28			HAI22 R2	Writing, creativity, and artificial intelligence. ChatGPT in the university context [13]	20 models	Academia	ideation item usage, creative thinking	T2T	University	originality score	Rule-based

Table 2. **Overview of studies included in the structured literature review on GenAI and creativity. The columns CP, CD, and HAI note the identification number of each study in the different meta-analyses across the different outcomes (CP: creative performance of human vs GenAI; CD: creative diversity in human-GenAI collaboration; HAI: creative performance in human-GenAI collaboration).**

3 RESULTS

We first summarize key characteristics of the included studies, and, afterward, we answer our research questions. Note that all analyses are reported at the effect size level (i.e., 127 observations with effect sizes for 28 studies).

3.1 Descriptive Summary

GenAI models: The capabilities of the GenAI model also determine how well it excels with creative tasks. In the included studies, *GPT-4* is the GenAI model that is used most frequently (37 of 127; 29.6%), followed by *GPT-3.5* (25, 20.0%). Other GenAI models appear in fewer than 11% of studies, which, on the one hand, reflects the state-of-the-art performance of GPT-4 in many tasks, but, on the other hand, implies that the findings may not generalize across all GenAI models but are subject to the specific choice (see our discussion in Section 4.3).

Task type: The majority of experiments involve task aimed at testing *creative writing* to assess creativity (48 of 127 tasks; 38%), followed by *creative problem solving* (25; 20%) and the alternate uses test (12, 10%). In contrast, only a few studies investigate business-oriented ideation tasks (fewer than 15%).

Participants: The choice of the participant sample determines whether findings later generalize only to laypeople or even to people with domain expertise. Here, the primary participant pool stems from an *Academic* background, accounting for over two-thirds of all observations (85 of 127; 67%), whereas *laypeople* (14) and business professionals (9) appear far less frequently. Of note, 19 studies do not even disclose the background of their participants or include participants pools with mixed backgrounds.

3.2 RQ1: Comparing the Creative Performance of Human vs. GenAI

In **RQ1**, we aim to answer: *How creative are ideas generated by GenAI compared to humans without GenAI support*, we first execute the random-effects model on the studies comparing (a) the creativity performance of humans versus GenAI. Here, out of the 127 observations, only 100 are relevant as they perform such a comparison.

Pooled effect: The pooled effect based on our random-effects meta-analysis corresponds to a Hedges $g = -0.048$ (95% CI: $[-0.257, 0.161]$; $p = 0.653$). The forest plot is shown in Figure 3. The overall heterogeneity is large ($I^2 = 98.90\%$; $\tau^2 = 1.08$). While the results hint toward a slight disadvantage toward GenAI, the effect sizes show no statistically significant difference between GenAI and human-only conditions across studies.

Robustness checks: Each study contributes less than 1.1 % weight, so the finding is not dominated by any single experiment. Nevertheless, we performed a leave-one-out sensitivity analysis. Here, the sequential deletion keeps the pooled estimate between $g = -0.083$ and $g = -0.024$, and with each 95% CI: still overlapping with zero and the I^2 staying above 98%. This shows that the results do not hinge on a single study but corroborate our general finding that there is no or only a negligible difference in the creative performance of humans vs. GenAI.
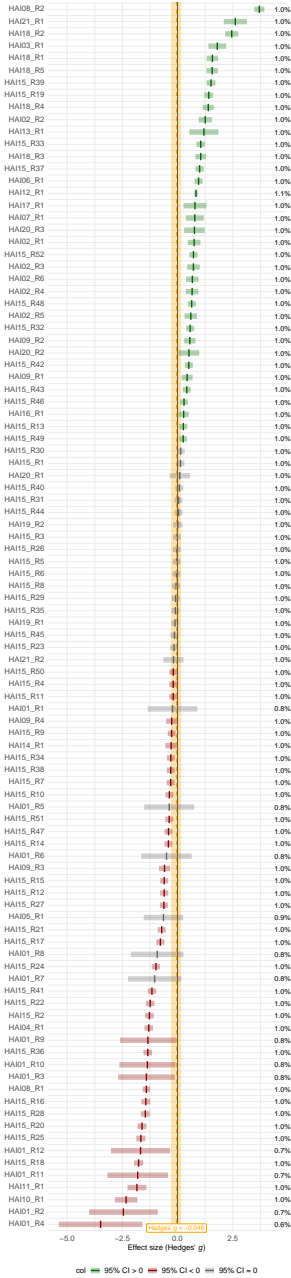
Fig. 3. **Pooled effect comparing the creative performance of humans vs. GenAI (RQ1).** The forest plot summarizes the Hedges' g effect sizes and 95% confidence intervals for a direct comparison between humans vs. GenAI (treatment: GenAI vs. control: human alone). Each line is one estimate (the weight is shown at the right), and the central diamond shows the overall effect size of g = -0.050, which indicates no statistically significant difference. The vertical line at g = 0 corresponds to a null effect; observations to the left favor the human control, and observations to the right favor GenAI.

(a) GenAI model.　　　　　　　(b) Participant type.　　　　　　　(c) Task type.
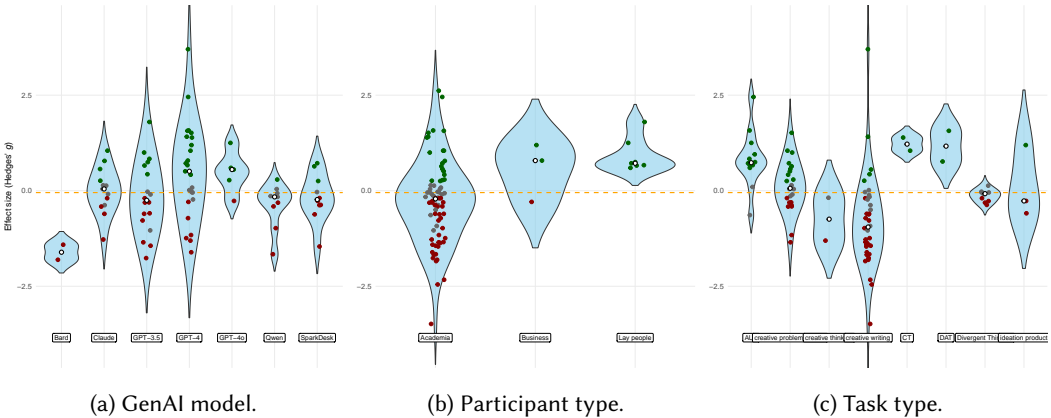
Fig. 4. **Heterogeneity analysis for RQ1 (creative performance of humans vs. GenAI).** Violin plots show the distribution of observation-level Hedges' g for the comparison in creative performance between human vs. GenAI conditions. The comparison is stratified by **(a)** GenAI model, **(b)** participant background, and **(c)** task type. Subgroup analyses are reported only for categories with a sufficient number of observations to support meaningful comparisons. The widths reflect the density of effect sizes; the dashed line corresponds to $g = 0$ with no overall difference.

3.2.1 Heterogeneity Analysis. **GenAI model (Figure 4a):** As expected, a key determinant for creative performance is the choice of the GenAI model. The model *GPT-4* performs the strongest, with distribution centered slightly above zero. A meta-regression confirms the positive effect for *GPT-4* ($g = 0.499$; 95% CI: $[0.132, 0.865]$; $p = 0.008$). For comparison, the effect for other models often overlaps with zero and has wide tails, which implies both non-significant coefficients and substantial within-model variability.

Participants (Figure 4b): For the moderator participants *laypeople* appear to be constantly outperformed by GenAI ($g = 0.918$; 95% CI: $[0.202, 1.630]$; $p = 0.012$). The distribution of laypeople is right-skewed and centered above zero. While academics show a small but significant pro-human difference ($g = -0.223$; 95% CI: $[-0.445, -0.001]$; $p = 0.049$), the results for participants with business background remain inconclusive ($g = 0.539$; 95% CI: $[-0.579, 1.660]$; $p = 0.345$). This suggests that the background of participants (and thus their domain expertise) is a moderator that explains the gap in creative performance between humans and GenAI.

Task type (Figure 4c): Violin shapes show modes above zero for AUT and CT tasks. The alternate uses test favors GenAI ($g = 0.855$; 95% CI: $[0.358, 1.350]$; $p < 0.001$), as do CT tasks ($g = 1.220$; 95% CI: $[0.016, 2.420]$; $p = 0.047$). By contrast, creative-writing tasks favor humans ($g = -0.717$; 95% CI: $[-1.010, -0.424]$; $p < 0.001$), a pronounced left-shift for creative writing. Other categories (e.g., creative problem solving, divergent thinking) are non-significant, highlighting task-specific effects. The violin plot has a broad overlap for the remaining categories, echoing the task-specific moderator coefficients.

3.2.2 Robustness. **Funnel plot & Egger test:** Egger's mixed-effects regression detects significant funnel asymmetry ($z = -3.363$; $p = 0.001$), indicating the presence of small-study or publication bias. The **trim-and-fill adjustment** therefore imputes 24 missing studies on the right of the funnel and shifts the random-effects estimate to ($g = 0.364$, 95% CI: $[0.131, 0.596]$; $p = 0.002$), implying that bias-correction would reverse the direction of the overall effect in favor of GenAI.

Influence diagnostics: All influence metrics (studentized residuals, DFFITS, Cook's D, hat values, cov.r) remain well below critical cut-offs, except for one flagged outlier; removing this study

lowers τ^2 only modestly (to 0.93) and leaves the pooled effect virtually unchanged, confirming that no single study drives the results.

3.3 RQ2a: Benefit of Human-GenAI Collaboration for Creative Performance

In **RQ2a**, we answer: *How creative are ideas generated by humans when supported by GenAI (compared to humans without GenAI support)?* Here, we restrict our analysis to 21 observations where studies report effect sizes aimed at understanding the effect of human-AI collaboration.

Pooled effect: The pooled effect based on our random-effects meta-analysis of amounts to Hedges $g = 0.273$ (95% CI: [0.018, 0.528]; $p = 0.036$). The forest plot is shown in Figure 5. Again, heterogeneity is substantial ($Q_{20} = 232.17$; $p < 0.001$; $I^2 = 94.11\,\%$; $\tau^2 = 0.3261$). This result indicates a small but statistically significant positive effect of GenAI support on the creativity of human-generated ideas, meaning that humans augmented with GenAI are more creative than humans without.

Robustness checks: No single study contributes more than 5.1 % weight, indicating that the positive performance advantage is not driven by any single experiment. Nevertheless, we performed a leave-one-out sensitivity analysis. Sequential deletion keeps the pooled effect between $g = 0.204$ and $g = 0.343$. In every iteration of the leave-one-out sensitivity analysis, the lower bound of the 95% CI stayed at or above 0.0004, and the I^2 remained above 91 %, confirming that the positive performance effect is stable yet accompanied by persistent heterogeneity.
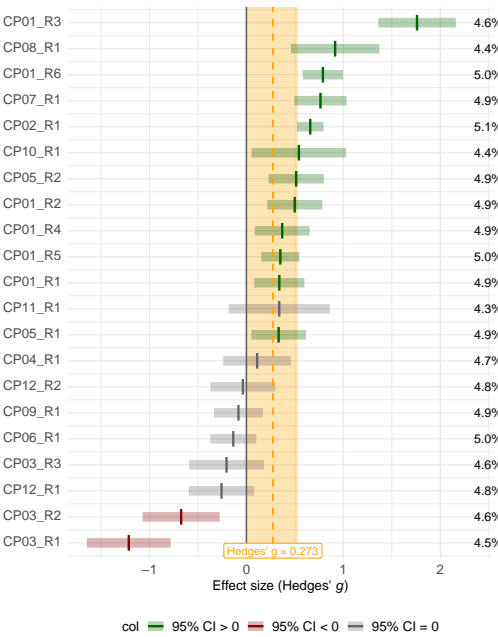
Fig. 5. **Pooled effect of the benefit from Human-GenAI collaboration on creative performance (RQ2a).** The forest plot summarizes the Hedges' g effect sizes and 95% confidence intervals (treatment: Human+GenAI collaboration versus control: human alone). Each line is one estimate (the weight is shown at the right), and the central diamond shows the overall effect size of $g = 0.219$, indicating a modest performance gain from GenAI assistance. The vertical line at $g = 0$ corresponds to a null effect; points to the right favor the GenAI-assisted collaboration.

(a) GenAI model. (b) Partcipant type. (c) Task type.
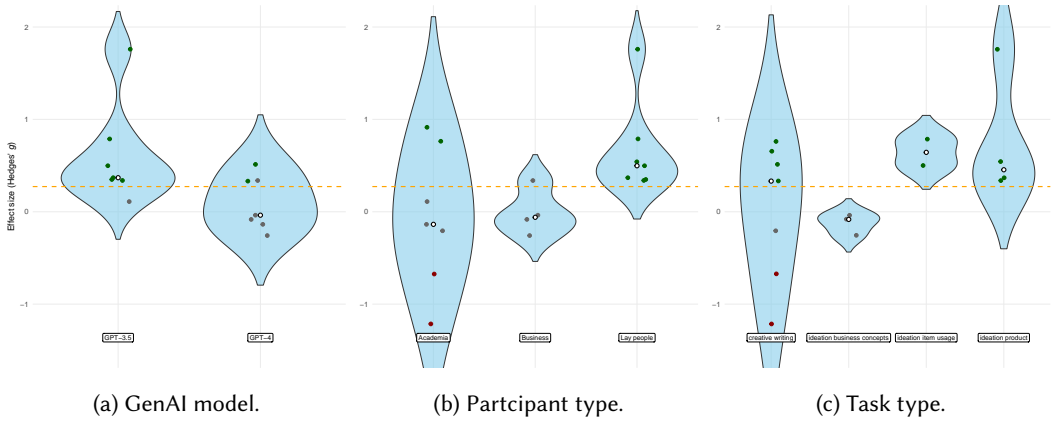
Fig. 6. **Heterogeneity analysis for RQ2a (creative performance of humans+GenAI vs. humans only).** Violin plots show the distribution of observation-level Hedges' g for the benefit in creative performance of human+GenAI collaboration over a human-only condition. The comparison is stratified by **(a)** GenAI model, **(b)** participant background, and **(c)** task type. Subgroup analyses are reported only for categories with a sufficient number of observations to support meaningful comparisons. The widths reflect the density of effect sizes; the dashed line corresponds to $g = 0$ with no overall difference.

3.3.1 Heterogeneity Analysis. **GenAI model (Figure 6a):** The plot shows a higher median and thicker right tail for *GPT-3.5* relative to *GPT-4*, which is also confirmed in a meta-analytic regression ($g = 0.587$; 95% CI: $[0.280, 0.893]$; $p < 0.001$). Interestingly, GPT-3.5 contributes larger performance gains, whereas *GPT-4* shows no detectable deviation from the overall mean ($g = 0.089$; 95% CI: $[-0.223, 0.401]$; $p = 0.577$).

Participants (Figure 6b): The distribution of effect sizes for *laypeople* is right-skewed and centered above zero, point to a benefit from collaboration ($g = 0.654$; 95% CI: $[0.237, 1.070]$; $p = 0.002$). The distribution of the effect sizes for *academia* ($g = -0.057$; 95% CI: $[-0.480, 0.367]$) and *business* ($g = -0.021$; 95% CI: $[-0.581, 0.540]$) cluster around the grand mean, indicating that participant expertise moderates outcomes asymmetrically. In other words, laypeople seem to benefit from co-creation with GenAI, while domain experts do not.

Task type (Figure 6c): A widest and mostly right-shifted violin is found for *ideation-product* tasks ($g = 0.743$; 95% CI: $[0.128, 1.360]$; $p = 0.018$) where human–GenAI teams show a better creative performance than the human-only condition. The effect sizes for tasks aimed at *creative writing* overlap with zero ($g = 0.048$; 95% CI: $[-0.412, 0.508]$). The same pattern is found for the other ideation tasks.

3.3.2 Robustness. **Funnel plot & Egger test:** Egger's regression detects no funnel asymmetry ($z = -0.624$; $p = 0.533$), suggesting the absence of small-study or publication bias. The **trim-and-fill procedure** imputes zero missing studies; the bias-adjusted estimate remains essentially unchanged at $g = 0.273$, ($95\%[0.018, 0.528]$), corroborating the robustness of the performance benefit. **Influence diagnostics:** One observation shows a high studentised residual ($r = 2.933$), but its removal reduces τ^2 only from 0.3261 to 0.2271 and leaves g within the original confidence limits. All other influence metrics (DFFITS, Cook's D, hat values, cov.r) remain below conventional thresholds.

3.4 RQ2b: Effect of Human-GenAI Collaboration on Creative Diversity

In **RQ2b**, we aim to answer: *How diverse are ideas generated by humans when supported by GenAI compared to humans without GenAI support.* Here, we analyze all studies comparing the creative performance of humans with vs. without GenAI support in regard to creative diversity. Note that creative diversity is measured only in a few studies (overall only 6 observations with effect sizes), because of which the meta-analytic results below are subject to a small sample size.

Pooled effect: As shown in Figure 7, the effect of human–AI collaboration on diversity yields a pooled Hedges' $g = -0.863$ ($95\%CI: [-1.328, -0.398], p < 0.001$). Still, heterogeneity is severe ($Q_5 = 51.69; p < 0.001; I^2 = 93.70\%; \tau^2 = 0.310$). Each study carries roughly 16% to 17% weight. Nevertheless, we find a consistent and statistically meaningful reduction in idea diversity when GenAI joins the team.

Robustness: We again performed a leave-one-out sensitivity analysis. Sequential deletion keeps the pooled estimate between $g = -0.655$ and $g = -0.952$; every 95% CI: excludes zero, and I^2 never falls below 78.21%. The diversity-reducing effect is therefore robust to the removal of any single study, albeit heterogeneity persists.
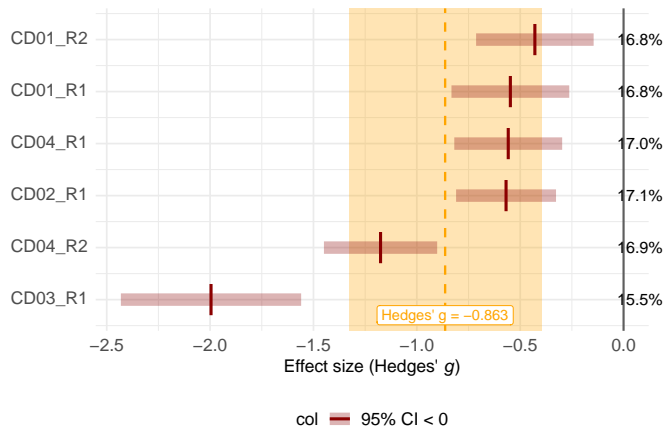
Fig. 7. **Pooled effects (RQ2b).** Forest plot summarizing the Hedges' g effect sizes and 95% confidence intervals for creative diversity at the observation level (treatment: human+GenAI vs. control: Human alone), across five individual replications (CD01_R1, CD02_R1, CD02_R2, CD03_R1, CD04_R1). Each line is one replication's estimate (weight at right), and the central diamond is the overall effect size of $g = 2.139$, indicating a substantial diversity gain from GenAI assistance. The vertical line at $g = 0$ marks the null; points to the right favor the GenAI-assisted treatment.

3.4.1 Heterogeneity Analysis. Due to the small sample size and limited variation in experimental designs across studies, we restrict our focus to the following subgroup analysis around participants (while we cannot perform subgroup analyses for GenAI models or task types as above).

Participants. Compared with laypeople samples, academic cohorts show a markedly stronger reduction in diversity ($g = -1.260$; 95% CI: $[-2.340, -0.187]$; $p = 0.021$), while business populations exhibit a non-significant shift ($g = -0.866$; 95% CI: $[-1.930, 0.197]$; $p = 0.110$). Participant background thus moderates the magnitude of the negative diversity effect.

3.4.2 Robustness. **Funnel plot & Egger test:** Eggers´ test and trimm-and-fill was not executed du to the amount of studies. **Influence diagnostics.** One outlier (study #4) shows a studentised residual

of -3.662 and Cook's $D = 0.766$, excising it lowers τ^2 from 0.310 to 0.067 and Q_E from 51.69 to 17.87, but the pooled effect remains negative and significant ($g = -0.656$; 95% CI: $[-0.913, -0.398]$). Thus, although the outlier inflates heterogeneity, it does not alter the substantive conclusion.

4 DISCUSSION

4.1 Summary of Key Findings

The fragmented state of the literature in human-GenAI co-creativity research raises two central questions: *How creative are the ideas generated by GenAI? And to what extent can GenAI support humans in producing ideas that are both creative and diverse?* For our meta-analysis, we screened 693 records and finally included 28 studies (with $N = 127$ subjects) comparing creative performance along three dimensions: (i) how creative GenAI is, (ii) how creative human+GenAI collaboration is vs. human-only performance, and (iii) the effect of human+GenAI collaboration on idea diversity.

RQ	Description	Hedges' g	Supported?
RQ1	Creative performance of GenAI (compared to human)	-0.05 (ns)	✗
RQ2a	Creative performance of human+GenAI (compared to human only)	0.27^*	✓
RQ2b	Diversity of ideas in human+GenAI (compared to human only)	-0.86^{***}	✓

Table 3. **Key findings from our meta-analysis.** Reported are Hedges' g for each anaysis. ✓ indicates $p < 0.001$, ✗ indicates non-significant. Significance levels: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

Finding for RQ1: Parity rather than super-human creativity. GenAI matches the average human creative output ($g \approx 0$; see Table 3). A moderate advantage ($g \approx 0.499$) emerges only for specific GenAI models and is largely attributable to a subset of studies using GPT-4. Hence, any observed superiority is tied to specific models, rather than being a general property of LLMs. This result is also in line with studies explicitly comparing the creative performance of different models, like ChatGPT-3 and ChatGPT-4 [37]. Nevertheless, with the growing number of research using recent LLMs with reasoning capabilities, we may observe a gradual shift in average performance benchmarks.

Finding to RQ2a. Modest but robust gains from collaboration. Human-GenAI teams deliver a small yet robust boost in creative performance ($g \approx 0.27$) that holds across different models, tasks, and participant backgrounds. While our meta-analysis focused on comparing human-AI collaboration with human-only performance (rather than contrasting human-GenAI teams with GenAI-only outputs), recent studies examining the latter report no significant performance gains from adding humans to GenAI-generated content [39]. Together, humans using GenAI therefore show higher creative performance than humans without GenAI support.

Finding to RQ2b. Decrease in idea diversity. Collaboration with GenAI is associated with a significant decrease in idea diversity (pooled $g \approx -0.86$), indicating a potential homogenization effect of the AI. This finding strengthens the impression, that collaboration with LLMs might improve individual performance while showing detrimental effects on group-level [17]. It remains unclear whether this effect generalizes beyond GPT-4 and text-based tasks, as this analysis was based on limited data.

Heterogeneity in our findings. Our heterogeneity analysis reveals three key insights. First, GenAI performs better on simple, standardized creativity tests—such as the alternative uses task (AUT) and consequences task (CT)—than on more complex, elaborative tasks like creative writing. Second, creative performance varies by model: newer models outperform older ones when generating ideas without human involvement, but, in human–AI collaboration, GPT-3.5 consistently

enhances creative output, while GPT-4 shows no significant added benefit. Third, laypeople particularly benefit from collaborating with GenAI, yet they tend to be outperformed when directly competing against it.

4.2 Implications

Practical implications: Our research suggests substantial potential for organizations to integrate GenAI into creative workflows—but only by pairing it with employees to enhance ideation processes. These benefits also imply a shift in individual creative potential: as GenAI can augment human creativity, especially lower innate creative capabilities might be compensated [17]. Yet, the observed gains in creativity could come at a cost, namely, in the form of a reduction in idea diversity. As a result, users collaborating with GenAI tend to produce more homogeneous outputs in some settings. This trade-off is especially problematic in contexts where ideational breadth is essential, such as where companies search for "out-of-the-box" ideas such as in open innovation and crowd ideation tasks [5]. Thus, while GenAI-human teaming may suffice for boosting individual creativity, practitioners will need to be careful when scaling GenAI use with the intention of promoting collective creativity.

Further, current GenAI models often tend to recombine familiar elements [39] across contexts rather than generating radically novel ideas, which may lead to more incremental innovation. This highlights the importance of human agency in co-creation processes to counteract potential convergence effects—where repeated exposure to similar outputs leads to reduced variability—and highlights the need for human-GenAI systems that actively support diversity.

Theoretical implications: Our study reveals that the creative impact of GenAI is highly contingent on contextual factors. A high between-study heterogeneity I^2 >80 % demonstrates that creativity augmentation cannot be treated as a general affordance of GenAI alone. However, the effectiveness mechanisms of GenAI in creativity remain unclear. Importantly, our heterogeneity analysis identifies only the choice of the GenAI model as a consistent moderator that explains higher creativity levels.

Given that the benefit from human-AI collaboration in creative tasks is robust for both laypeople and domain experts, it is likely that the performance gain is due to a generic facilitation mechanism (e.g. faster drafting and broader ideation, reduced cognitive load) rather than task-specific benefits from augmentation. Previous evidence indicates that co-creation tasks are particularly beneficial for human-AI collaboration when there are inherent synergies between humans and GenAI systems, allowing each to leverage their respective strengths—unlike in many analytical, decision-making contexts [35].

4.3 Limitations and Future Research

Our meta-analysis has limitations due to the current state of research and technological progress. First, some of the included studies are not yet peer-reviewed. Nevertheless, we still chose to include them to provide a timely synthesis of empirical findings including recent ones. Second, the findings depend on the choice of GenAI models. Given the fast-evolving capabilities of GenAI, it is likely that newer models, including different reasoning capabilities or different fine-tuning strategies, may yield different outcomes. Third, many studies use simple, well-known tasks for benchmarking such as the so-called alternative uses test, which asks participants to generate creative uses for common objects. Such tasks may have been part of the training data in GenAI models, which is a known issue in GenAI research [61] and underscores the need for devising novel tasks that preserve experimental rigor while minimizing overlap with model training data.

To guide future work, we identified salient gaps around the use of GenAI for creativity systems in HCI/CSCW research based on our structured literature review and propose three key research directions.

Research direction 1: *More relevant, real-world-inspired study designs:* The majority of studies focus on simplified creativity tasks such as the alternative uses test (AUT) [69]. While these tasks draw on established scales from psychological research to assess creativity, they fall short of capturing the complexity of real-world settings that involve creative thinking. In contrast, there are only a few studies that test GenAI in real-world situations such as work settings (e.g., [10]), or with creative tasks beyond text, such as images, audio, or code.

Future studies could focus on more realistic scenarios for a higher ecological validity and further explore context-sensitive moderators for more targeted creativity interventions. For example, to enhance relevance to CSCW and HCI applications in business contexts, future studies could investigate settings where tacit knowledge plays a critical role, or where creative tasks must account for specific organizational contexts—such as operational constraints or strategic alignment. One promising direction would be to examine how GenAI can support ideation in tasks like developing new business models that align with a company's brand strategy or existing capabilities of that company. Moreover, all of the existing studies were performed at a single time point; thus, longitudinal designs are needed to examine how collaboration patterns and creativity effects evolve over time—especially as users gain fluency in leveraging GenAI within creative tasks.

Research direction 2: *Understand psychological mechanisms:* Existing research has primarily focused on the outcomes of GenAI-assisted creativity, typically through rigorous A/B experiments. Yet, prior research has largely overlooked the underlying psychological mechanisms that drive the effects. For example, it is unclear whether gains in GenAI-assisted creativity come from broader ideation (e.g., generating more ideas due to increased speed or reduced distractions), deeper engagement (e.g., more intensive thinking due to reduced cognitive load), or heightened motivation (e.g., playful interaction with the system).

To develop a more nuanced understanding of human-GenAI co-creation, future work should explore potential psychological mediators and moderators, such as cognitive effort, user agency, trust, task framing, and participant's personality traits. Uncovering such cognitive processes is relevant to identifying how human and machine capabilities can be combined to foster synergistic collaboration. Ultimately, such insights could inform frameworks with a notion of 'appropriate reliance' (see [53] for an overview on the notion of appropriate reliance in AI advice), that is, when it is most beneficial for an idea to originate from the human, the GenAI system, or through their interaction.

Research direction 3: *Explore CSCW/HCI design choices to elicit creative thinking:* The majority of studies identified in our literature review rely on relatively naive GenAI setups—often based on simple one-shot prompts (e.g., as in [37]) or by providing participants with unmodified, off-the-shelf LLMs (e.g., as in [39]). In contrast, studies that devise and compare different design choices are largely absent. Future studies could thus explore more elaborate workflows (e.g., critique-refine, suggestions-refine) that reflect an iterative creation process. Eventually, this could help to integrate GenAI tools into creativity systems through effective, user-friendly interfaces that allow for natural interactions, as well as to build GenAI agents that are effective at supporting creative thinking tasks.

5 CONCLUSION

Our meta-analysis shows that GenAI and humans show similar creative performance on average, with moderate advantages emerging mainly for GPT-4. In contrast, human-AI collaboration shows small but consistent gains in creative output across tasks and contexts. However, collaboration

with GenAI reduces the diversity of ideas, indicating a risk of creative outputs that could become more homogeneous. Our meta-analysis also highlights important gaps in the literature, which provide interesting avenues for future research to the understand psychological mechanisms of human-GenAI augmentation and identify drivers for how GenAI system can successfully elicit creative thinking.

REFERENCES

[1] Anja Eisenreich, Julian Just, Daniela Gimenez-Jimenez, and Johann Füller. 2024. Revolution or inflated expectations? Exploring the impact of generative AI on ideation in a practical sustainability context. *Technovation* 138 (2024), 103123. doi:10.1016/j.technovation.2024.103123

[2] Jaan Aru. 2025. Artificial intelligence and the internal processes of creativity. *The Journal of Creative Behavior* (2025). doi:10.1002/jocb.1530

[3] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How AI ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. https://arxiv.org/abs/2401.13481

[4] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to meta-analysis*. John Wiley & Sons, Ltd, West Sussex, UK.

[5] Kevin J. Boudreau and Karim R. Lakhani. 2013. Using the crowd as an innovation partner. *Harvard Business Review* 91, 4 (2013), 60–69.

[6] Léonard Boussioux, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani. 2024. The crowdless future? Generative AI and creative problem-solving. *Organization Science* 35, 5 (2024), 1589–1607.

[7] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics* 140, 2 (2025), 889–942. doi:10.1093/qje/qjae044

[8] Noah Castelo, Zsolt Katona, Peiyao Li, and Miklos Sarvary. 2024. How AI outperforms humans at creative idea generation. *Available at SSRN 4751779* (2024). doi:10.2139/ssrn.4751779

[9] Gary Charness and Daniela Grieco. 2024. Creativity and AI. *Available at SSRN 4686415* (2024). doi:10.2139/ssrn.4686415

[10] Zenan Chen and Jason Chan. 2024. Large language model in creative work: The role of collaboration modality and user expertise. *Management Science* 70, 12 (2024), 9101–9117. doi:10.1287/mnsc.2023.03014

[11] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge, New York. doi:10.4324/9780203771587

[12] Zheyuan Kevin Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2024. The effects of generative ai on high skilled work: Evidence from three field experiments with software developers. *Available at SSRN 4945566* (2024). doi:10.2139/ssrn.4945566

[13] de Vicente-Yagüe-Jara, María-Isabel, Olivia López-Martínez, Verónica Navarro-Navarro, and Francisco Cuéllar-Santiago. 2023. Writing, creativity, and artificial intelligence: ChatGPT in the university context. *Comunicar: Media Education Research Journal* 31, 77 (2023), 45–54.

[14] Jonathan J. Deeks, Julian P. T. Higgins, Douglas G. Altman, and on behalf of the Cochrane Statistical Methods Group. 2019. Analysing data and undertaking meta-analyses: 10. In *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd, 241–284. doi:10.1002/9781119536604.ch10

[15] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7, 3 (1986), 177–188.

[16] Anil R. Doshi, Sen Chai, and Matthias Troebinger. 2024. How experience moderates the impact of generative AI ideas on the research process. *Available at SSRN 5013086* (2024). doi:10.2139/ssrn.5013086

[17] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (2024), eadn5290. doi:10.1126/sciadv.adn5290

[18] William P. Dunlap, Jose M. Cortina, Joel B. Vaslow, and Michael J. Burke. 1996. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods* 1, 2 (1996), 170.

[19] Sue Duval and Richard Tweedie. 2000. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56, 2 (2000), 455–463.

[20] Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *bmj* 315, 7109 (1997), 629–634.

[21] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative AI. *Business & Information Systems Engineering* 66, 1 (2024), 111–126. doi:10.1007/s12599-023-00834-7

[22] Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E. Robertson, Steve Rathje, Jochen Hartmann, Saif M. Mohammad, Oded Netzer, Alexandra A. Siegel, Barbara Plank, and Jay J. van Bavel. 2025. Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology* 4, 2 (2025), 96–111. doi:10.1038/s44159-024-00392-z

[23] Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. https://arxiv.org/abs/2310.08433v1

[24] Simone Grassini and Mika and Koivisto. 2025. Artificial creativity? Evaluating AI against human performance in creative interpretation of visual stimuli. *International Journal of Human–Computer Interaction* 41, 7 (2025), 4037–4048. doi:10.1080/10447318.2024.2345430

[25] Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S. Christian, John Patrick, Laura Huang, Yoed N. Kenett, and Kevin Lewis. 2019. "Forward flow": A new measure to quantify free thought and predict creativity.

         American Psychologist 74, 5 (2019), 539–554.

[26]   Matthew Grimes, Georg von Krogh, Stefan Feuerriegel, Floor Rink, and Marc Gruber. 2023. From Scarcity to Abundance: Scholars and Scholarship in an Age of Generative Artificial Intelligence. *Academy of Management Journal* 66, 6 (2023), 1617–1624. doi:10.5465/amj.2023.4006

[27]   Joy Paul Guilford, Paul R. Christensen, Philip R. Merrifield, and Robert C. Wilson. 1978. Alternate uses. (1978). doi:10.1037/t06443-000

[28]   Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports* 13, 1 (2023), 18617. doi:10.1038/s41598-023-45644-9

[29]   Kent F. Hubert, Kim N. Awa, and Darya L. Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (2024), 3440. doi:10.1038/s41598-024-53303-w

[30]   Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024. Creativity in AI: Progresses and challenges. https://arxiv.org/abs/2410.17218v4

[31]   Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. https://arxiv.org/abs/2411.02316v5

[32]   Dan Jackson and Jack Bowden. 2016. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails? *BMC Medical Research Methodology* 16, 1 (2016), 118. doi:10.1186/s12874-016-0219-y

[33]   Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences* 118, 25 (2021), e2022340118. doi:10.1073/pnas.2022340118

[34]   Jennifer Haase and Paul H.P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity* 33, 3 (2023), 100066. doi:10.1016/j.yjoc.2023.100066

[35]   Jonathan A. Plucker, Ronald A. Beghetto, and Gayle T. Dow and. 2004. Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist* 39, 2 (2004), 83–96. doi:10.1207/s15326985ep3902{_}1

[36]   Anuj Kapoor and Madhav Kumar. 2025. Frontiers: Generative AI and personalized video advertisements. *Marketing Science* (2025). doi:10.1287/mksc.2023.0494

[37]   Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports* 13, 1 (2023), 13601. doi:10.1038/s41598-023-40858-3

[38]   Larry V. Hedges. 1981. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6, 2 (1981), 107–128. doi:10.3102/10769986006002107

[39]   Byung Cheol Lee and Jaeyeon Chung. 2024. An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour* 8, 10 (2024), 1906–1914. doi:10.1038/s41562-024-01953-1

[40]   Mark A. Runco and Garrett J. Jaeger and. 2012. The standard definition of creativity. *Creativity Research Journal* 24, 1 (2012), 92–96. doi:10.1080/10400419.2012.650092

[41]   Jack McGuire, David de Cremer, and Tim van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports* 14, 1 (2024), 18525. doi:10.1038/s41598-024-69423-2

[42]   Peidong Mei, Deborah N. Brewis, Fortune Nwaiwu, Deshan Sumanathilaka, Fernando Alva-Manchego, and Joanna Demaree-Cotton. 2025. If ChatGPT can do it, where is my creativity? Generative AI boosts performance but diminishes experience in creative writing. *Computers in Human Behavior: Artificial Humans* 4 (2025), 100140. doi:10.1016/j.chbah.2025.100140

[43]   Noah Bohren, Rustamdjan Hakimov, and Rafael Lalive. 2024. Creative and strategic capabilities of Generative AI: Evidence from large-scale experiments: IZA Discussion Papers. https://hdl.handle.net/10419/305744

[44]   William Orwig, Emma R. Edenbaum, Joshua D. Greene, and Daniel L. Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior* 58, 1 (2024), 128–136. doi:10.1002/jocb.636

[45]   Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sally E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. PRISMA 2020 Checklist. https://www.prisma-statement.org/prisma-2020-checklist

[46]   Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sally E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher.

966 2021. PRISMA 2020 Flow Diagram. https://www.prisma-statement.org/prisma-2020-flow-diagram

[47] Satu Parjanen. 2012. Experiencing creativity in the organization: From individual creativity to collective creativity. *Interdisciplinary Journal of Information, Knowledge & Management* 7 (2012), 109–128. doi:10.28945/1580

[48] Robert A. Peterson and Steven P. Brown. 2005. On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology* 90, 1 (2005), 175.

[49] Ralph L. Rosnow and Robert Rosenthal. 1996. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods* 1, 4 (1996), 331.

[50] Suebsarn Ruksakulpiwat, Lalipat Phianhasin, Chitchanok Benjasirisan, Kedong Ding, Anuoluwapo Ajibade, Ayanesh Kumar, and Cassie Stewart. 2024. Assessing the efficacy of ChatGPT versus human researchers in identifying relevant studies on mHealth interventions for improving medication adherence in patients with ischemic stroke when conducting systematic reviews: Comparative analysis. *JMIR mHealth and uHealth* 12 (2024), e51526. doi:10.2196/51526

[51] Solve Sæbø and Helge Brovold. 2024. On the stochastics of human and artificial creativity. https://arxiv.org/abs/2403.06996v1

[52] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. doi:10.1145/3514094.3534128

[53] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *International Conference on Intelligent User Interfaces (IUI)*. doi:10.1145/3581641.3584066

[54] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. https://arxiv.org/abs/2409.04109

[55] Yu Song, Longchao Huang, Lanqin Zheng, Mengya Fan, and Zehao Liu. 2025. Interactions with generative AI chatbots: unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving. *International Journal of Educational Technology in Higher Education* 22, 1 (2025), 12. doi:10.1186/s41239-025-00508-2

[56] Liz Sanders SonicRim. 2001. Collective creativity. *Design* 6, 3 (2001), 1–6.

[57] Jonathan A. C. Sterne, Jelena Savović, Matthew J. Page, Roy G. Elbers, Natalie S. Blencowe, Isabelle Boutron, Christopher J. Cates, Hung-Yuan Cheng, Mark S. Corbett, Sandra M. Eldridge, Jonathan R. Emberson, Miguel A. Hernán, Sally Hopewell, Asbjørn Hróbjartsson, Daniela R. Junqueira, Peter Jüni, Jamie J. Kirkham, Toby Lasserson, Tianjing Li, Alexandra McAleenan, Barnaby C. Reeves, Sasha Shepperd, Ian Shrier, Lesley A. Stewart, Kate Tilling, Ian R. White, Penny F. Whiting, and Julian P. T. Higgins. 2019. RoB 2: a revised tool for assessing risk of bias in randomised trials. *bmj* 366 (2019), 4898. doi:10.1136/bmj.l4898

[58] Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. 2024. Large language models show both individual and collective creativity comparable to humans. https://arxiv.org/abs/2412.03151

[59] Yuan Sun, Eunchae Jang, Fenglong Ma, and Ting Wang. 2024. Generative AI in the wild: Prospects, challenges, and strategies. In *CHI Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3613904.3642160

[60] Susannah B.F. Paletz and Kaiping Peng. 2008. Implicit theories of creativity across cultures: Novelty and appropriateness in two product domains. *Journal of Cross-Cultural Psychology* 39, 3 (2008), 286–302. doi:10.1177/0022022108315112

[61] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[62] Wolfgang Viechtbauer. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36, 3 (2010), 1–48. doi:10.18637/jss.v036.i03

[63] Wolfgang Viechtbauer and Mike W.-L. Cheung. 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* 1, 2 (2010), 112–125. doi:10.1002/jrsm.11

[64] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The effects of generative AI on design fixation and divergent thinking. In *CHI Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3613904.3642919

[65] Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J. Su, Zhun Deng, Michael Qizhe Xie, Hannah Brown, and Kenji Kawaguchi. 2024. Can AI be as creative as humans? https://arxiv.org/abs/2401.01623v4

[66] Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across LLMs. https://arxiv.org/abs/2501.19361v1

[67] Robert C. Wilson, Joy P. Guilford, and Paul R. Christensen. 1953. The measurement of individual differences in originality. *Psychological Bulletin* 50, 5 (1953), 362.

[68] Zhikun Wu, Thomas Weber, and Florian Müller. 2025. One does not simply meme alone: Evaluating co-creativity between LLMs and humans in the generation of humor. In *International Conference on Intelligent User Interfaces (IUI)*. doi:10.1145/3708359.3712094

[69] Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. https://arxiv.org/abs/2406.04244v1

[70] Man Zhang, Ying Li, Yang Peng, Yijia Sun, Wenxin Guo, Huiqing Hu, Shi Chen, and Qingbai Zhao. 2025. AI delivers creative output but struggles with thinking processes. https://arxiv.org/abs/2503.23327v1

[71] Jiexin Zheng, Ka Chau Wong, Jiali Zhou, and Tat Koon Koh. 2024. Large language model in ideation for product innovation: An exploratory comparative study. *Available at SSRN 4729982* (2024). doi:10.2139/ssrn.4729982

[72] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (2024), pgae052. doi:10.1093/pnasnexus/pgae052

[73] Wenbo Zou and Feng Zhu. 2025. Generative AI adoption in human creative tasks: Experimental evidence. *Available at SSRN 5196748* (2025). doi:10.2139/ssrn.5196748

Acknowledgment of AI Usage