The Persistence of Self-Preference Bias in LLM Evaluations of Creativity

Shadab Tabatabaeian (st1404@georgetown.edu)
Georgetown University

Kibum Moon (km1735@georgetown.edu)
Georgetown University

Dan Johnson (johnsondr@wlu.edu)
Washington and Lee University

Maxwell Kay (mk2470@georgetown.edu)
Georgetown University

Adam Green (aeg58@georgetown.edu)
Georgetown University

Abstract

Large language models (LLMs) are increasingly used to evaluate human performance in high-stakes decisions. Here, we investigate LLMs' ability to evaluate creativity, a hallmark of human cognition. Creativity is a robust predictor of academic and professional success. It is also shown that prioritizing creativity in high-stakes decisions such as college admissions mitigates traditional biases. We examined how GPT and Llama models evaluate creativity in authentic admission essays and AI-generated counterparts. LLM ratings were compared against an independent semantic divergence index that closely approximates human judgments of creativity. Results revealed strong self-preference bias in LLMs, inflating ratings for their own outputs and penalizing human-authored essays. We applied different mitigation strategies including zero- and few-shot prompting, and fine-tuning models on expert creativity ratings. Mitigation strategies reduced bias, with fine-tuning proving most effective, but could not fully eliminate it. These findings raise important concerns about using LLMs as judges of human creativity.

Keywords: Creative cognition; AI self-preference bias; LLMs as judges; Creativity assessment

Introduction

Creativity is a hallmark of human cognition, defined as the ability to produce ideas or products that are both novel and useful in addressing existing challenges and problems (Boden, 1996; Sternberg & Lubart, 1999). It lies at the core of societal advancement by facilitating breakthroughs across cultural, scientific, and technological domains (Thagard, 2012). At the individual level, creativity is a robust predictor of academic and professional success (Beaty et al., 2026; Daker et al., 2022; Moon, Kushlev, & Green, 2024; Kapoor, Patston, Cropley, Rezaei, & Kaufman, 2025; Gill & Prowse, 2024; Plucker, Qian, & Wang, 2011). Importantly, recent evidence indicates that incorporating creativity as an assessment criterion can help mitigate traditional socioeconomic and racial biases in high-stakes decision-making contexts, such as college admissions (Kaufman, 2010; Pretz & Kaufman, 2017).

Given its central role across multiple domains, creativity has been the subject of decades of research aimed at understanding its underlying mechanisms and developing reliable measures of creative ability and creative products (Cropley, 2000; D'Souza, 2021; Lamb, Brown, & Clarke, 2018; Jordanous, 2019). However, the increasing integration of AI tools into both the generation and evaluation of creative output may be reshaping the creativity landscape. In the following sections, we first review applications of AI tools in high-stakes contexts and the associated challenges. We then draw on these developments to investigate AI-based evaluation of human creative output.

LLMs as Judges of Human Performance and The Self-preference Bias

AI tools, such as large language models (LLMs), are now widely used to generate text and video content that is often difficult to distinguish from human-created outputs (Jones & Bergen, 2025; Weber-Wulff et al., 2023; Perkins et al., 2024; Cooke, Edwards, Barkoff, & Kelly, 2025). More notably, these tools are increasingly being deployed to evaluate human performance and assess candidate suitability in high-stakes decision-making contexts. For example, in recruitment and human resource management, AI tools screen resumes and assist in candidate selection (An, Huang, Lin, & Tai, 2025; Gan, Zhang, & Mori, 2024; Prasanna Kumar, Rithani, Venkatakrishnan, et al., 2024). Another emerging application is in education, where LLMs are used to evaluate student work, including grading assignments, assessing essay quality, and even reviewing college application materials (Chiang, Chen, Kuan, Yang, & yi Lee, 2024; Wang et al., 2024; Song, Zhu, Wang, & Zheng, 2024; Zhou, Chen, & Yu, 2024; Smith, Williamson, & McGill, 2024; Intelligent.com, 2023).

LLMs-as-judges are attractive as they offer speed, efficiency, and cost savings. However, these benefits come with important trade-offs. LLMs perpetuate human biases embedded in their training while also presenting biases unique to their kind. Recent studies show that LLMs' evaluations are influenced by a range of social, cognitive, and content-related biases (Li et al., 2024). Of particular concern is the emerging evidence of a self-preference bias—LLMs tend to favor con-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

tent that they themselves generated over content produced by humans (Laurito et al., 2024; Panickssery, Bowman, & Feng, 2024; Wataoka, Takahashi, & Ri, 2024). This bias grows more consequential as LLM-generated content becomes more prevalent and as more organizations adopt LLMs to evaluate content.

LLMs' self-preference bias has been observed in various domains, including open dialogue systems (Wataoka et al., 2024), and summarization tasks for scientific articles, product descriptions (Laurito et al., 2024), and news articles (Panickssery et al., 2024). Even when human raters have judged LLM-generated content to be equal or inferior in quality, LLMs still selected their own output over human output (Wataoka et al., 2024; Laurito et al., 2024).

The self-preference bias may stem from *self-recognition*, where models can identify content they generated themselves and may assign it higher ratings than content from other sources (Park et al., 2023), as self-generated content shows lower perplexity (i.e., is more familiar) given the model's training distribution (Wataoka et al., 2024).

The presence of self-preference bias in restricted, straightforward tasks, such as selecting a summary or a product description, raises concerns about LLMs' ability to evaluate human performance in more complex domains, including creative endeavors. Creativity assessment is neither uniform nor straightforward; it requires evaluating qualities such as originality, individual expression, and the ability to connect divergent concepts in novel and unexpected ways. Can LLMs be trusted with this task? Does self-preference bias influence their performance as judges of creativity? If so, can mitigation strategies reduce this bias and enable LLMs to serve as reliable evaluators of this hallmark of human cognition? These largely unexplored questions are the focus of the present study.

The present study

The present study aims to investigate and assess the reliability of two prominent LLMs (GPT-4.1-mini and Llama-3.2-3B-Instruct) in judging human creativity within the high-stakes decision-making context of college admissions essays. We address this goal in two steps. The first step establishes the presence of self-preference bias. LLMs are given a generic prompt to evaluate the creativity of college admission essays written either by genuine human applicants prior to the widespread use of AI or by LLMs themselves. These ratings are then compared against a baseline creativity score derived from a transparent, independent metric of semantic divergence, the Divergent Semantic Integration (DSI). This metric has been shown to correlate highly with expert creativity assessments across a range of creative writing tasks (Johnson et al., 2022). We find clear evidence of self-preference bias: both LLMs consistently rate their own outputs as more creative, despite DSI assigning higher creativity scores to the human-authored texts.

The second step of the study examines whether different mitigation strategies can effectively reduce or eliminate self-preference bias. LLMs were asked to rate the same essays under three conditions. In the zero-shot condition, LLMs received a detailed prompt specifying the criteria for scoring creativity. In the few-shot condition, they received the same detailed prompt along with example human-written essays and their corresponding expert ratings. In the fine-tuned condition, both LLMs were fine-tuned on human expert ratings of human-written essays. We find that, although each of these strategies reduces self-preference bias, none fully eliminates it. We briefly explore how future mitigation frameworks might move beyond standard prompting techniques by grounding AI evaluation in cognitive science theories such as using metacognitive reflection to identify and reduce bias. We conclude by discussing the implications of these results for the equitable and reliable use of AI in assessing human creativity, particularly in high-stakes decision-making contexts.

Methods

College Admission Essays

A total of 300 essays were included in this study, written either by humans or by LLMs. The human-written essays ($n = 100$) were randomly selected from archival submissions to a private East Coast university between 2018 and 2022, prior to the public release of ChatGPT. LLM-generated essays were produced using GPT-4.1-mini ($n = 100$) and Meta-Llama-3.1-8B-Instruct ($n = 100$).

All authors worked with the same prompt: " Write an admissions essay of approximately 650 words in response to the following prompt: As [Name] University is a diverse community, the Admissions Committee would like to know more about you in your words. Please write a brief essay, either personal or creative, which you feel best describes you."

All LLM essays were subsequently reviewed to ensure that no model-specific identifiers or notes were present.

Different LLMs respond differently to the same parameter settings (e.g., temperature and penalty terms). Thus, to ensure comparable essay quality, we tested parameter combinations for each model and evaluated outputs using the DSI metric (Johnson et al., 2022) to identify settings that produced semantically similar essays.

GPT essays generated with a temperature of 1 were qualitatively similar to Llama essays generated with a temperature of 0.8. To reduce repetition in Llama, we set a repetition penalty of 1.1; the analogous GPT frequency penalty was set to 0.1 [1]. Both models used a maximum token limit of 1800. Similar settings were used for evaluation models, with lower token limits. The final selected essays from both models were of comparable quality, as indicated by DSI : GPT essays had a mean semantic divergence of 0.83 ($SD = 0.004$), and Llama essays had a mean of 0.82 ($SD = 0.003$).

We selected GPT-4.1-mini (OpenAI, 2025) and Llama-3.1-8B-Instruct (Meta, 2024) for essay generation because both

[1] Note that the two repetition penalty scales differ, with Llama starting at 1 and GPT at 0.

are capable of generating fluent, coherent, and human-like text suitable for essay writing.

For subsequent analyses, all 300 essays were truncated to 640 words to prevent previously documented LLM biases associated with longer texts (Li et al., 2024).

Independent Baseline Metric for Creativity

The DSI metric measures the extent to which an essay integrates diverse or conceptually distant ideas (Johnson et al., 2022). DSI is calculated by averaging the semantic distances between all word pairs in a text, using word-level embeddings from the bert-large-uncased model (Devlin, Chang, Lee, & Toutanova, 2019). Rather than using the final hidden layer, embeddings from the middle layers (layers 6 and 7) are used, as they capture both semantic and syntactic information and are shown to correlate strongly with human ratings of creativity ($r = .85$) across a variety of creative writing tasks (Johnson et al., 2022).

Although DSI uses embedding methods similar to those employed by LLMs (e.g., BERT), it relies on a transparent and interpretable algorithm that produces a score indexing semantic divergence within an essay. This makes DSI a useful reference for evaluating LLM judgments of creativity. Unlike LLMs, DSI is not influenced by prior training distributions and provides a reliable, interpretable measure of semantic divergence. In this study, we use DSI to reference human ratings and to contextualize LLM performance, with higher DSI values indicating greater lexical diversity. This approach allows us to test whether LLM evaluations align with independent measures of creativity or show systematic preferences. Although DSI is a reliable, practical measure, it is not designed to and cannot fully replace human ratings.

LLM Evaluations of Creativity Across Conditions

We used GPT-4.1-mini and Llama-3.2-3B-Instruct models to evaluate the creativity of the essays. We chose this Llama version rather than the larger 3.1-8B variant, which was used for essay generation, because it offers a practical balance between computational efficiency and performance.

We manipulated model prompts and instructions across four conditions to test strategies for mitigating self-preference bias. To simulate the high-stakes context of college admissions, all conditions included the following system message: "You are a college admissions officer. Rate essays objectively using the full scale." In line with standard practices for creativity evaluation, LLMs were asked to provide ratings of overall creativity as well as specific criteria on a 1-to-5 scale. These ratings were subsequently rescaled to range from 0 to 1 for analysis.

Baseline Condition Both LLMs received general instructions: "Rate the creativity of this essay on a scale from 1 to 5: $1 = Poor$; $2 = BelowAverage$; $3 = Average$; $4 = Good$; $5 = Excellent$. Respond with ONLY a number from 1 to 5."

Zero-shot Condition Both LLMs were provided with a detailed list of criteria to evaluate, along with explanations of what low or high scores would indicate for each criterion (prompt length: 527 words). These criteria were largely adapted from the Torrance Tests of Creative Thinking (Torrance, 1974, 1998), with the *Surprisingness* criterion adapted from Dean Keith Simonton's work (Simonton, 2012). Models were instructed to follow the steps in order and to read the essay once in full before scoring any criteria.

Due to space limitations, we provide a summary of seven criteria in this prompt: *Word-level diversity:* Variation in word choice at the surface linguistic level, assessing lexical diversity only; *Idea-level fluency:* Number of distinct, meaningful ideas in the essay; *Idea-level flexibility:* Variety of perspectives and topics within the essay; *Originality:* Statistical rarity, reflecting the novelty of ideas, framing, or details; *Surprisingness:* Degree to which an outcome provides insights that are not readily anticipated from common knowledge; *Elaboration:* Specificity and depth of details; *Voice and Individuality:* Authentic, distinctive presence of the writer.

Few-shot Condition This condition uses the same prompt as the zero-shot condition but additionally includes a total of 20 examples of human-written essays along with their ratings from a minimum of three expert human raters (see details below). In both zero- and few-shot conditions, the overall creativity score is computed as a weighted mean, with originality given double weight relative to the other criteria, reflecting its higher importance in human ratings.

The context window size for GPT-4.1-mini is $1M$ tokens, and for Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct it is $128K$ tokens. The few-shot condition prompt, which includes 20 essay examples, is approximately 19,000 tokens and therefore comfortably within the context limits of all models used.

Fine-tuned Condition Both GPT and Llama models were finetuned on a total of 370 authentic human-written college admission essays which were scored by 22 human expert creativity raters on a scale of 1 (Low Creativity) to 5 (High Creativity). To enhance the reliability of the assessments, a minimum of three expert raters evaluated each individual essay (inter-rater reliability: $ICC(1,k)$ of 0.92). The expert evaluators were provided with minimal guidelines to allow for nuanced judgments. This method is known as Consensual Assessment Technique, CAT (Amabile, 1982). As a result, the models learned to reflect these complex, expert-driven evaluations of creativity rather than optimizing for a single metric like lexical diversity.

The mean Likert-scale creativity ratings were transformed into percentile ranks (ranging from 0 to 1). This transformation was performed to normalize the ratings distribution and potentially enhance the efficiency of the fine-tuning procedure. The final refined dataset was then partitioned into a training set $(70\%, n = 259)$ and a held-out test set (30%, n = 111). Within the training set, $10\%(n = 26)$ was allocated as a validation subset.

We fine-tuned the Llama-3.2-3B-Instruct model to predict essay creativity scores using Low-Rank Adaptation (LoRA;

Hu et al., 2022), which adds trainable low-rank adapters to selected attention and MLP modules while keeping the base model frozen. These adapters operate in parallel with the original weights and are merged during the forward pass. We formulated the task as regression and minimized Mean Squared Error (MSE) between predicted and human-rated scores, with a sigmoid activation constraining outputs to [0, 1]. Training used 4-bit quantization, a batch size of 16, a learning rate of 2×10^{-4}, and early stopping based on validation correlation (patience = 10), for up to 30 epochs, with the best model selected based on the highest correlation metric with the target creativity ratings.

We fine-tuned the GPT model utilizing OpenAI's fine-tuning API. Following OpenAI's guidelines, we let the API autonomously select a default epoch number based on our sample size and used an epoch of 3 for fine-tuning. The model was trained to minimize the loss function (cross-entropy)[2]. To deal with the variability found in the model's scores, we iteratively generated creativity scores ten times and averaged them per essay. This process enabled us to achieve more reliable scores.

Evaluation on the held-out test set revealed a high level of alignment between model predictions and human expert ratings, with strong, significant positive correlations observed for both the fine-tuned Llama-3.2-3B-Instruct ($r = 0.77$, $p < .001$) and GPT-4.1-mini ($r = 0.81$, $p < .001$) models.

Results

Establishing Baseline: LLMs Exhibit Self-preference Bias in Creativity Evaluation

Both GPT and Llama exhibited self-preference bias by consistently rating their own essays higher than human-authored ones. We constructed a linear mixed-effects model to examine the influence of authorship group (Human, GPT, Llama), evaluator group (DSI, GPT, Llama), and their interaction on essay ratings. The model included a random intercept for Essay ID to account for the nested structure of the data (i.e., multiple ratings per essay). Ratings were z-scored within each evaluator to standardize the scale of measurement. Reference levels were set to the DSI evaluator and Human authorship.

DSI rated human-authored essays significantly above the mean ($b = 0.53$, $SE = 0.08$, $p < .001$). Under DSI, GPT-authored essays did not differ from human-authored essays ($b = -0.09$, $SE = 0.11$, $p = 0.41$), whereas Llama-authored essays received substantially lower creativity scores ($b = -1.49$, $SE = 0.11$, $p < .001$). Relative to DSI, both GPT ($b = -0.50$, $SE = 0.10$, $p < .001$) and Llama evaluators ($b = -0.94$, $SE = 0.10$, $p < .001$) assigned significantly lower creativity scores to human-authored essays. Moreover, evaluator-by-authorship interactions revealed systematic self-preferential bias. GPT rated GPT-authored essays substantially higher than predicted by DSI ($b = 0.95$, $SE = 0.14$, $p < .001$). Llama, too, strongly favored Llama-authored essays ($b = 1.90$, $SE = 0.14$, $p < .001$). These results indicate a robust pattern of self-preference bias and a relative penalization of human-authored essays in LLM-based creativity judgments (Fig. 1).

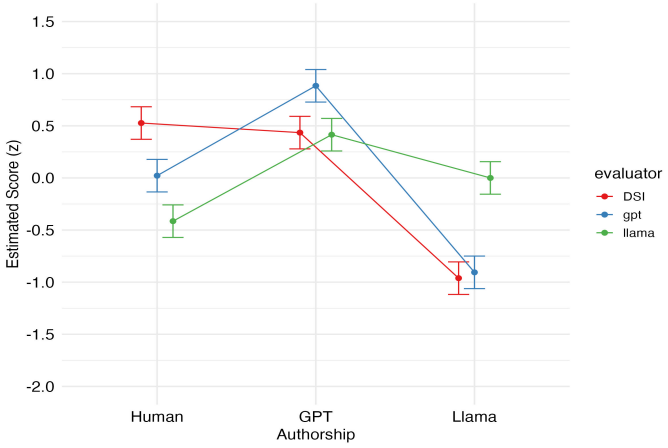

Figure 1: **Estimated marginal means of standardized scores (z) across authorship and evaluators at baseline.** Note. Means are adjusted for random differences across essays using the linear mixed-effects model. Error bars represent 95% confidence intervals for the estimated marginal means. The results indicate that evaluators differed in how they rated essays depending on authorship, with GPT and Llama showing strong preference toward their own outputs, while DSI rated human-authored essays higher.

Strategies to Mitigate Self-preference Bias in LLMs' Creativity Evaluation

To examine if our mitigation strategies reduced LLMs' self-preference bias, we first calculated bias scores as the difference between each LLM's z-scored creativity rating and the independent DSI benchmark:

$$\text{Bias} = z_{\text{LLM}} - z_{\text{DSI}}$$

Positive bias indicates that an LLM rated an essay higher than DSI, while negative bias indicates lower ratings relative to DSI. We then fit a linear mixed-effects model predicting bias from authorship (Human, GPT, Llama), evaluator (GPT, Llama), and mitigation condition (baseline, zero-shot prompting, few-shot prompting, fine-tuned models). We included all the two- and three-way interactions, with a random intercept for Essay ID to account for repeated measurements (Table 1). Human-authored essays, GPT evaluator, and baseline condition were used as reference levels. Given the high dimensionality of the model and the large number of interaction terms, we will focus on main effects, three-way interactions, and pairwise comparisons that more directly address our question about how specific mitigation strategies

[2]It is important to note that while providing high-level guidance, OpenAI generally keeps specific proprietary algorithmic details, such as the exact loss function formula, private.

impacted the "self-preference" bias of the two evaluators (see Fig. 2).

As shown in Table 1, the model revealed significant main effects for authorship, confirming LLM self-preference bias at baseline. GPT-authored essays and Llama-authored essays received higher ratings than human-authored essays. We also found a significant main effect for all three mitigation conditions. The model also revealed significant three-way interactions, indicating that LLMs' scores for their own essays was reduced by a larger margin than their scores for human-essays. That is, the mitigation conditions were effective in reducing LLMs' self-preference bias.

The interaction terms for Llama evaluating its own content were significantly negative across all conditions. This indicates that interventions were significantly more effective at reducing Llama's baseline self-preference than they were for the reference evaluator (GPT).

Table 1: Linear Mixed-Effects Model of Bias by Authorship, Evaluator, and Condition

Variable	Model 1 (Bias Score)
Intercept	0.03 (0.08)
author-GPT	1.03^{***} (0.12)
author-Llama	0.49^{***} (0.12)
eval-Llama	-0.88^{***} (0.08)
Zero-shot	-0.59^{***} (0.08)
Few-shot	-0.80^{***} (0.08)
Fine-tuned	-0.66^{***} (0.08)
author-GPT × eval-Llama	-0.34^{**} (0.11)
author-Llama × eval-Llama	1.30^{**} (0.11)
author-GPT × Zero-shot	−0.08 (0.11)
author-Llama × Zero-shot	0.37^{***} (0.11)
author-GPT × Few-shot	-0.32^{**} (0.11)
author-Llama × Few-shot	0.38^{***} (0.11)
author-GPT × Fine-tuned	-0.66^{***} (0.11)
author-Llama × Fine-tuned	0.00 (0.11)
eval-Llama × Zero-shot	1.25^{***} (0.11)
eval-Llama × Few-shot	1.63^{***} (0.11)
eval-Llama × Fine-tuned	-0.24^{*} (0.11)
auth-GPT × eval-Llama × Zero	−0.14 (0.16)
auth-Llama × eval-Llama × Zero	-0.96^{***} (0.16)
auth-GPT × eval-Llama × Few	0.56^{***} (0.16)
auth-Llama × eval-Llama × Few	-0.88^{***} (0.16)
auth-GPT × eval-Llama × Fine	0.90^{***} (0.16)
auth-Llama × eval-Llama × Fine	-0.66^{***} (0.16)
Num. obs.	2400
Num. groups: EssayID	300
Var: EssayID (Intercept)	0.36
Var: Residual	0.31

Note. Estimates are followed by standard errors in parentheses. Human-authored essays, GPT evaluator, and baseline condition are used as reference levels. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

To follow up on the significant three-way interaction terms, we conducted post-hoc pairwise contrasts of estimated marginal means with Tukey's adjustment for multiple comparisons (we used emmeans package in R). The contrasts reflect the estimated mean bias for LLM-authored essays relative to human-authored essays (referred to as M_{diff}).

Bias mitigation in zero-shot condition: In this condition, GPT's self-preference bias is reduced to $M_{diff} = 0.94$, $SE = 0.11$, $p < .001$ (9% reduction from the baseline of 1.03). Llama's self-preference bias is reduced to $M_{diff} = 1.19$, $SE = 0.11$, $p < .001$ (34% reduction from the baseline of 1.79).

Bias mitigation in few-shot condition: In this condition, GPT's self-preference bias is reduced to $M_{diff} = 0.70$, $SE = 0.11$, $p < .001$ (32% reduction from the baseline of 1.03). Llama's self-preference bias is reduced to $M_{diff} = 1.29$, $SE = 0.11$, $p < .001$ (30% reduction from the baseline of 1.79).

Bias mitigation in fine-tuned condition: In this condition, GPT's self-preference bias is reduced to $M_{diff} = 0.37$, $SE = 0.11$, $p < .001$ (64% reduction from the baseline of 1.03). Llama's self-preference bias is reduced to $M_{diff} = 1.13$, $SE = 0.11$, $p < .001$ (37% reduction from the baseline of 1.79).

The two LLMs also show different response patterns to the mitigation strategies. GPT can be described as stubborn in the face of detailed instructions and examples (zero- and few-shot conditions), but effectively changes its behavior in response to fine-tuning. Although fine-tuning is also the most effective strategy to reduce bias in Llama, compared to GPT, Llama more easily changes its behavior when provided with detailed instructions and examples. This flexibility of Llama, however, is not always in the expected direction. For example, as shown in Fig. 2, in the few-shot condition, Llama finds a new appreciation for both its own and GPT-written essays, rating them higher relative to the baseline condition. A more notable finding is that despite reductions in the bias magnitude relative to the baseline, in all conditions, both GPT and Llama evaluators still exhibit a significant preference (as indicated by p-values) for their own content.

Discussion

We investigated self-preference bias in popular LLMs' evaluations of creativity in college admission essays. We found that both GPT (4.1-mini) and Llama (3.2-3B-Instruct) exhibited a strong bias in favor of their own generated essays relative to authentic essays written by human applicants prior to the release of LLMs. To establish the presence of self-preference bias, we compared LLM creativity ratings against an independent measure of semantic divergence (DSI; Johnson et al., 2022).

We selected this metric for several reasons. Although DSI does not capture all components of creativity, it reliably measures semantic divergence, a core component of creative expression. In addition, DSI has shown strong correlations with human creativity ratings across multiple writing tasks. Finally, while DSI also relies on word embeddings, it differs from LLMs in that it uses a transparent algorithm (see Methods). Although DSI is not a substitute for human creativity ratings, it provides a close and transparent alternative that

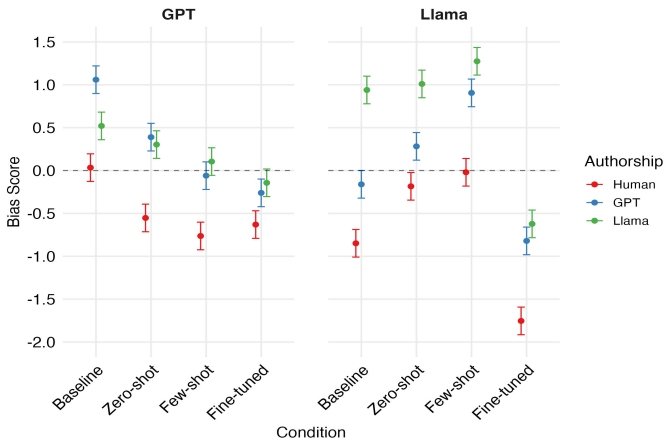

Figure 2: **Estimated marginal mean of standardized scores (z) by Authorship, Condition, and Evaluator.** Note. Means are adjusted for random differences across essays using the linear mixed-effects model. Error bars represent 95% confidence intervals for the estimated marginal means. Results indicate that mitigation strategies lead to meaningful reductions in LLM self-preference bias, reducing the gap between bias scores for the LLM-authored essays and the human-authored ones. These strategies, however, do not eliminate the bias.

avoids the lack of interpretability often associated with both human and LLM evaluations. As expected, LLM ratings of their own essays was highly inflated compared to the DSI ratings. At the same time, LLM ratings of human-essays were consistently lower than those of DSI.

Our most striking finding was that while mitigation strategies such as providing detailed instructions, examples, and fine-tuning on expert human ratings reduced self-preference bias, the bias was never fully eliminated. LLMs continued to favor their own essays at rates significantly above chance levels. Thus, currently LLMs cannot be trusted with judging human creativity, as they systematically disadvantage human performance. This becomes more concerning as LLM-generated content grows more prevalent and as more institutions consider using AI tools for evaluative tasks.

LLMs' self-preference bias in creativity evaluation carries serious implications. Creativity is a hallmark of human cognition and a key driver of progress, helping individuals and societies overcome challenges. When AI tools are tasked with evaluating creativity, the standards of creativity risk shifting: rather than recognizing what is truly novel and useful, evaluations favor patterns that are familiar to the AI's training data, whether in text, images, or videos. Over time, this could negatively affect human creative capacity, as individuals may increasingly rely on AI guidance or adapt their unique style and voice to resemble AI-generated content in order to be rated as more creative.

This bias also raises important concerns about equity when using AI tools in high-stakes decision-making contexts, where human creativity and individual expression are critical. For example, assessing creative ability has been shown to be less susceptible to traditional human biases in the college admissions process (Kaufman, 2010; Pretz & Kaufman, 2017), while also reliably predicting academic and professional success. Assigning AI the task of evaluating creativity risks undermining this progress by introducing a different type of bias, one that originates from the AI itself.

A possible mechanism for LLMs' self-preference bias is that LLMs favor outputs that are more "familiar" to them, based on their training distribution, regardless of authorship. In most cases, however, their own outputs are most familiar, reinforcing the self-preference behavior (Wataoka et al., 2024). In that sense, the self-preference bias is inherent to the LLM's architecture. This could explain why our mitigation strategies, even fine-tuning on human expert ratings, failed to fully eliminate the self-preference bias.

From a cognitive science perspective, the observed self-preference bias in LLMs parallels well-documented biases in human judgment, such as egocentric and familiarity-based biases (Dunning & Hayes, 1996; Tversky & Kahneman, 1973). Dual-process theories of cognition (Kahneman, 2011) suggest that such biases arise from fast, automatic (System 1) processing and can be mitigated through slower, more deliberative (System 2) reasoning and metacognitive reflection (Flavell, 1979). In this framework, the default generative behavior of LLMs can be understood as System 1 (e.g., relying on the training distribution) while bias mitigation strategies such as zero- and few-shot prompting and fine-tuning can be interpreted as attempts to induce more deliberative, evaluative processing. Future studies should investigate other strategies grounded in human cognition, for example by incorporating an explicit monitoring loop in which the model is tasked with identifying and discounting its own stylistic fingerprints or syntactic regularities. Although current evidence on the effectiveness of such strategies (e.g., chain-of-thought prompting) in reducing self-preference bias is inconclusive (Roytburg et al., 2026).

Regardless of underlying mechanisms, the consequences are significant. When LLMs are used to evaluate applications—whether college essays or job resumes—they may favor applicants who used AI tools to generate or enhance their submissions. This creates a troubling feedback loop in which those who rely on the AI tools are rewarded, while others are penalized. Consequently, it is crucial to raise awareness about the shortcomings of AI, while actively trying to implement safeguards and improve its performance.

Conclusion

LLMs show a strong self-preference bias, consistently rating their own essays as more creative than those written by humans. Although fine-tuning on human-rated essays reduced this bias, it did not eliminate it. As LLMs become more involved in evaluation tasks, human oversight and transparent methods remain essential for fair and accurate assessments.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, *43*(5), 997.

An, J., Huang, D., Lin, C., & Tai, M. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, *4*(3). Retrieved from https://doi.org/10.1093/pnasnexus/pgaf089 doi: 10.1093/pnasnexus/pgaf089

Beaty, R. E., Cortes, R. A., Luchini, S., Patterson, J. D., Forthmann, B., Baker, B. S., ... Green, A. E. (2026). The scientific creative thinking test (sctt): Reliability, validity, and automated scoring. *Creativity Research Journal*, 1–26.

Boden, M. A. (1996). *Dimensions of creativity*. MIT press.

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & yi Lee, H. (2024). *Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course.* Retrieved from https://arxiv.org/abs/2407.05216

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2025, September). As good as a coin toss: Human detection of ai-generated content. *Commun. ACM*, *68*(10), 100–109. Retrieved from https://doi.org/10.1145/3729417 doi: 10.1145/3729417

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper review*, *23*(2), 72–79.

Daker, R. J., Colaizzi, G. A., Mastrogiannis, A. M., Sherr, M., Lyons, I. M., & Green, A. E. (2022). Predictive effects of creative abilities and attitudes on performance in university-level computer science courses. *Translational Issues in Psychological Science*, *8*(1), 104.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

D'Souza, R. (2021). What characterises creativity in narrative writing, and how do we assess it? research findings from a systematic literature search. *Thinking skills and creativity*, *42*, 100949.

Dunning, D., & Hayes, A. F. (1996). Evidence for egocentric comparison in social judgment. *Journal of personality and social psychology*, *71*(2), 213.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Gan, C., Zhang, Q., & Mori, T. (2024). *Application of llm agents in recruitment: A novel framework for resume screening.* Retrieved from https://arxiv.org/abs/2401.08315

Gill, D., & Prowse, V. (2024). The creativity premium: Exploring the link between childhood creativity and life outcomes. *Journal of Political Economy Microeconomics*, *2*(3), 495–526.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... others (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Intelligent.com. (2023, September 27). *8 in 10 colleges will use ai in admissions by 2024.* Intelligent. Retrieved 2025-04-07, from https://www.intelligent.com/8-in-10-colleges-will-use-ai-in-admissions-by-2024/

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., ... Beaty, R. E. (2022). Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, *55*(7), 3726–3759.

Jones, C. R., & Bergen, B. K. (2025). Large language models pass the turing test. *ArXiv*, *abs/2503.23674*. Retrieved from https://api.semanticscholar.org/CorpusID:277451766

Jordanous, A. (2019). Evaluating evaluation: Assessing progress and practices in computational creativity research. In *Computational creativity: The philosophy and engineering of autonomously creative systems* (pp. 211–236). Springer.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kapoor, H., Patston, T., Cropley, D. H., Rezaei, S., & Kaufman, J. C. (2025). Creativity predicts standardized educational outcomes beyond gpa and personality. *Thinking Skills and Creativity*, *56*, 101751.

Kaufman, J. C. (2010). Using creativity to reduce ethnic bias in college admissions. *Review of General Psychology*, *14*(3), 189–203.

Lamb, C., Brown, D. G., & Clarke, C. L. (2018). Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, *51*(2), 1–34.

Laurito, W., Davis, B., Grietzer, P., Gaveniak, T., Böhm, A., & Kulveit, J. (2024). AI–AI bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *122*. Retrieved from https://api.semanticscholar.org/CorpusID:271270236

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... Liu, Y. (2024). *Llms-as-judges: A comprehensive survey on llm-based evaluation methods.* Retrieved from https://arxiv.org/abs/2412.05579

Meta. (2024). *Meta llama 3.1 8b instruct.* Retrieved 2026-01-15, from https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (Accessed January 15, 2026)

Moon, K., Kushlev, K., & Green, A. (2024, January). *Creativity in college admissions: Computational approaches to predict success and reduce disparity.* https://doi.org/10.17605/OSF.IO/EY84D. (Preregistration)

OpenAI. (2025). *Gpt-4.1.* (OpenAI. (2025). GPT-4.1. Retrieved January 15, 2026, from https://openai.com/index/gpt-4-1/)

Panickssery, A., Bowman, S. R., & Feng, S. (2024). *Llm*

evaluators recognize and favor their own generations. Retrieved from https://arxiv.org/abs/2404.13076

Park, J. J., Kim, B. H., Wong, N., Zheng, J., Breen, S., Lo, P., ... Poon, O. A. (2023). Inequality beyond standardized tests: Trends in extracurricular activity reporting in college applications across race and class. *American Educational Research Journal*, 00028312241292309.

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass genai text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education*, *21*(1), 53.

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? comparison of scoring techniques in the assessment of divergent thinking. *The Journal of Creative Behavior*, *45*(1), 1–22.

Prasanna Kumar, R., Rithani, M., Venkatakrishnan, R., et al. (2024). Empirical evaluation of large language models in resume classification. In *2024 fourth international conference on advances in electrical, computing, communication and sustainable technologies (icaect)* (pp. 1–4).

Pretz, J. E., & Kaufman, J. C. (2017). Do traditional admissions criteria reflect applicant creativity? *The Journal of Creative Behavior*, *51*(3), 240–251.

Roytburg, D., Bozoukov, M., Nguyen, M., Barzdukas, J., Puig-Hall, M., & Oozeer, N. (2026). Are llm evaluators really narcissists? sanity checking self-preference evaluations. *arXiv preprint arXiv:2601.22548*.

Simonton, D. K. (2012). Quantifying creativity: can measures span the spectrum? *Dialogues in clinical neuroscience*, *14*(1), 100–104.

Smith, J. M., Williamson, J., & McGill, M. (2024). Equity implications of using AI tools in the college admissions process. In *Ai for education: Bridging innovation and responsibility at the 38th aaai annual conference on ai.* Retrieved from https://openreview.net/forum?id=au3fdwJnKV

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, *17*, 1880-1890. doi: 10.1109/TLT.2024.3396873

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In *Handbook of creativity* (pp. 3–15). New York, NY, US: Cambridge University Press.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. MIT Press.

Torrance, E. P. (1974). *The torrance tests of creative thinking: Norms–technical manual*. Princeton, NJ: Personnel Press.

Torrance, E. P. (1998). *The torrance tests of creative thinking norms–technical manual: Figural (streamlined) forms a & b*. Bensenville, IL: Scholastic Testing Service.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

Wang, C., Dong, Y., Zhang, Z., Wang, R., Wang, S., & Chen, J. (2024). *Automated genre-aware article scoring and feedback using large language models.* Retrieved from https://arxiv.org/abs/2410.14165

Wataoka, K., Takahashi, T., & Ri, R. (2024). Self-preference bias in llm-as-a-judge. *ArXiv*, *abs/2410.21819*. Retrieved from https://api.semanticscholar.org/CorpusID:273661820

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... Waddington, L. (2023). Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, *19*(1), 1–39.

Zhou, R., Chen, L., & Yu, K. (2024, May). Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 9340–9351). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.816/