TWO STUDIES ON ASSESSING AI-AUGMENTED CREATIVITY WITH LARGE LANGUAGE MODELS

By

Jiaoping Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Business Administration – Information Technology Management – Doctor of Philosophy

2025

ABSTRACT

Large Language Models (LLMs) have been increasingly integrated into a variety of tasks, facilitating human endeavors in generating creative outputs, ranging from product ideation to digital artwork. Such novel capabilities of LLMs have ushered in a new era of collaboration between humans and Artificial Intelligence (AI), which has grabbed the attention of researchers and practitioners alike. Thus, in this dissertation, I explore the intersection of emerging LLMs and creativity, with a primary focus on writing tasks. This dissertation includes two studies. In the first study, I examine the impact on perceived creativity of varying levels of generative capabilities of LLMs - namely, randomness, which has been overlooked so far and which is manipulated via a quasi-experiment. I find that collaborating with an LLM with high randomness that generates more diverse advice does not necessarily lead to increased perceived creativity of work, as the role of humans matters. Moreover, I explore how the characteristics of human evaluators and their perceived extent of AI use influence their assessments of creativity. In the second study, I focus on growing concerns regarding the potential misuse of generative AI, particularly its capacity to produce plagiarized content. Motivated by the divergent thinking creativity literature using the Divergent Association Task (DAT), I construct *DAT(Sent)*, a metric to proxy semantic dissimilarities within a document, and further propose an effective GPT detector classifier, *GPT-DATector*. I show that on average, human-generated contents have a larger *DAT(Sent)* than AI-generated texts across different writing tasks and datasets. Empirical evaluations demonstrate that the proposed *GPT-DATector* outperforms state-of-the-art models in terms of prediction performance. Most importantly, *GPT-DATector* has the potential to reduce bias in the detection of AI-generated text.

ACKNOWLEDGMENTS

I am sincerely grateful to my dissertation chair, Professor Anjana Susarla, whose invaluable support and encouragement have guided me throughout this journey. She always believes in my abilities and gives me the confidence to persevere. Her example as a scholar—committed to continual learning, impactful research, and constant contributions to the academic community—has left a lasting influence on me.

I would like to express my gratitude to my committee members, Professor Chenhui Guo, Professor Quan Zhang, and Professor Musaib Ashraf, for their insightful feedback and generous time. Their guidance has significantly strengthened this dissertation. I also extend my appreciation to Professor Laura Brandimarte and Professor Uttara M. Ananthakrishnan for their mentorship and support throughout my academic path.

It has been an honor to be part of the Eli Broad College of Business at Michigan State University over the past six years. I am grateful for the financial support provided by the Department of Accounting and Information Systems at the Eli Broad College of Business. I also wish to thank the Center for Ethical and Socially Responsible Leadership (CESRL) from Michigan State University, for the research funding that supported the data collection and cloud-based resources essential to this dissertation.

Moreover, I want to extend my appreciation to the faculty, staff, my fellow Krishna Pothugunta, and my office mate Li Zhang from the AIS department. Their support has been important throughout my time in the program. I am especially thankful to my lunchmates, Shuting Wu, Tom Shang, and Sangmok Lee, for their companionship and friendship, which brought encouragement and laughter along the way. I am also deeply grateful to all my friends who have offered their support throughout my Ph.D. studies.

Most importantly, I owe my deepest thanks to my parents, Yuying Ou and Chaoran Chen. Their unconditional love and support have been the foundation of my personal and academic growth. They gave me the freedom to explore, taught me to take responsibility for my choices, and instilled in me the importance of staying positive and empathetic. Their guidance has shaped how I see the world—with curiosity, openness, and care.

Lastly, thank my family, Di Wu and Jackson Wu, who have given me immense love and countless cherished memories throughout my Ph.D. journey. They have stood by me during the challenges and celebrated with me during the milestones, making this journey even more meaningful. Their presence in my life keeps reminding me of the warmth of love and family.

These years of pursuing my Ph.D. have been a journey of personal growth. It has taught me how to rise after each setback, how to navigate moments of solitude, and how to learn quickly and master critical knowledge. Throughout this process, I have constantly shifted between vulnerability and resilience, at times feeling discouraged by stagnant research, while at other times gaining renewed confidence in myself. This journey has required constant self-reflection and honest conversations with myself, helping me recalibrate and continue forward.

In the end, despite the challenges, I made it through. Reflecting on these years, I feel more resilient, more open-minded, and more committed to lifelong learning than when I began. I will forever be grateful for my time in Michigan, and it will always hold a special place in my heart.

TABLE OF CONTENTS

CHAPTER 1 Do Large Language Models' Generative Capabilities Boost Creativity? Assessing AI-Augmented Creativity With LLMs

1.1 Introduction

The advent of large-scale generative language models (LLMs), trained on vast quantities of data, has adapted to a wide range of downstream applications such as bar exams (Katz et al., 2024) and academic assignments (Stokel-Walker, 2022). Along with the impact of LLMs in enhancing performance on objective, well-defined tasks, a more complex and underexplored domain is understanding their transformative role in human-AI collaborations for subjective, open-ended tasks. One of such domains is the creativity industries, which encourages more novel and unexpected content. Prior research has begun to explore facilitating human endeavors in generating creative outputs, ranging from product ideation to problem-solving (Wang, Yang, and Sun, 2023; Boussioux et al., 2024). However, less attention has been paid to narrative writing, arguably one of the activities that best represents and distinguishes human intelligence (Arriagada, 2020). Therefore, in this study, we mainly focus on human-AI collaboration in the context of writing creative narratives, particularly examing how the three elements, AI artifacts, human creators, and human-AI collaboration modes, affect the perceived creativity of work.

The literature discusses how adopting AI artifacts (e.g., LLM adoption) affects work performance, mostly by perceiving LLMs as static assistants using the default mode (Fügener et al., 2022; Chen and Chan, 2024). For instance, prior studies examine the extent of creative output produced by AI alone or human-AI collaborations (Wang, Yang, and Sun, 2023; Zhou and Lee, 2024). However, as generative models, LLMs possess a key capability: generating more or less divergent outputs for a given input. Such generative capability can be systematically modulated through the adjustment of model parameters (such as the temperature parameter), thereby designing AI assistants with differentiated capabilities. Despite growing interest in addressing the importance of harnessing LLMs' generative capabilities to power human-AI interaction designs (Wang et al., 2021; Lee, Liang, and Yang, 2022), limited empirical evidence focuses on how various LLMs' generative capabilities influence the perceived creativity of co-produced work. Specifically, in the context of creative writing, to what extent does the LLMs' capability to generate more or less diverse responses affect the perceived creativity of the work? Therefore, to comprehensively explore the impact of AI artifacts with varying LLMs' inherent generative capabilities, particularly randomness, we raise **RQ1(a)**: *In the human-AI collaboration process, how do varying generative capabilities of LLMs affect the perceived creativity of work as evaluated by external evaluators?*

Importantly, while some people appreciate AI's potential to enhance human productivity (Logg, Minson, and Moore, 2019), others may not be as welcoming to this new type of collaborators, as one may feel threatened by their capabilities as humans, resulting in distrust in the output they produce and in what some refer to as algorithm aversion (Dietvorst, Simmons, and Massey, 2018). This aversion is revealed as a hesitation to accept and use advice from algorithms, despite their advice often being fairly good (Kleinmuntz, 1990). In particular, little is known about the impact of humans' algorithm aversion during human-AI collaboration (Brynjolfsson, 2023). Thus, except for the impact of AI artifacts (e.g., with varying inherent generative capabilities), we further explore the moderating role of human collaborators' algorithm aversion (e.g., the extent to which human creators utilize LLMs' advice). Accordingly, we ask **RQ1(b)**: *In the human-AI collaboration process, how does human creators' algorithm aversion moderate the effect of LLMs' generative capabilities on* *perceived creativity of work?*

Next, we turn our attention to how variations in human-AI collaboration patterns shape creative outcomes. While prior research has primarily drawn on human–IT collaboration and algorithm interactions (Bauer, von Zahn, and Hinz, 2023; Revilla et al., 2023; Shaikh and Vaast, 2023), empirical investigations have largely focused on outcome-based evaluations such as task performance (Zhou and Lee, 2024; Vaccaro, Almaatouq, and Malone, 2024) and user experience (Jakesch et al., 2023; Mirowski et al., 2023; Bauer, von Zahn, and Hinz, 2023). Yet, as AI systems exhibit increasingly generative and interactive capacities that resemble human behavior, questions emerge regarding the applicability of frameworks rooted in human–human collaboration (Goffman, 2017). Specifically, the interaction order literature emphasizes that the social situational factors (e.g., autonomy) dominate over individual social structures (e.g., demographics like race). Extending this perspective to human–AI collaboration requires capturing evolving interactional patterns across varying contexts. To this end, we draw on the literature on human-human dyadic collaboration (Damon and Phelps, 1989; Storch, 2002), which identifies four distinct interaction patterns based on levels of equality (i.e., balanced contribution) and mutuality (i.e., interactive engagement). By substituting one human collaborator with an LLM, we seek to explore **RQ1(c)**: *In the human-AI collaboration process, for a given level of LLMs' generative capabilities and algorithm aversion among human creators, which interaction pattern(s) result in the highest perceived creativity of the work?*

Last, human-AI interactions can alter the ways in which opinions are formed Jakesch et al. (2023), such as human-involved evaluation processes. As more users engage with LLMs in daily tasks, it is unclear how these experiences may shift what people consider creative. Despite human-AI co-creation processes, in this study, we also evaluate the human evaluation process by investigating the effect of two individual characteristics: 1) human evaluators' familiarity with LLMs, and 2) their perception of AI use when evaluating work without knowing its origin. Thus, we ask **RQ2**: *Do human evaluators' experience with LLMs and perceived AI use by writers affect the perceived creativity of the work?*

To answer these questions, we employ a structured experimental framework by focusing on creative writing tasks under two distinct conditions—low vs high randomness—in the human-AI collaboration process. Randomness is one of the main characteristics that affect LLMs' capabilities: high randomness allows LLMs to generate more diverse advice, while low randomness tends to cause LLMs to generate less varied and more repetitive suggestions. We further design and conduct an online survey to assess the perceived creativity of co-generated narratives by external evaluators.

Our findings provide significant insights into human-AI interaction by considering both collaboration and external evaluation processes. From the human-AI collaboration perspective, we find that simply collaborating with LLMs with greater generative capabilities (i.e., high randomness LLM where AI generates more diverse suggestions) does not lead to increased perceived creativity of work compared to collaborating with low-randomness LLMs. But the moderating role of human collaborators really matters. In high-randomness settings, reduced algorithm aversion, through greater utilization of AI-generated advice, is associated with higher perceived creativity of outputs. Conversely, in low-randomness settings (characterized by less diverse AI recommendations), diminished algorithm aversion correlates with lower perceived creativity. When it comes to varying human-AI collaboration patterns, we find that a human creator with low algorithm aversions, combined with high mutuality and human-led collaboration, is associated with the highest perceived creativity in work. From the human evaluation perspective, we find human evaluators who are more experienced with LLMs or perceive greater AI use (even though the nature of writers is not disclosed) tend to consider narratives as more creative.

We contribute to the growing literature related to human-AI augmented creativity by studying the consequences of collaborating with LLMs with varying inherent generative capabilities. Such findings have implications for efficient deployments of AI assistance in work involving creative tasks, which complement existing research on the synergistic impact of AI and humans (Kane et al., 2021; Chen and Chan, 2024; Boussioux et al., 2024). We also contribute to the human-AI collaboration literature by building up theoretical frameworks motivated by the human-human collaboration literature (Damon and Phelps, 1989; Storch, 2002). We show that greater interactions between human collaborators and AI, particularly when human-led, lead to higher perceived creativity in co-created work. Moreover, our study's findings provide additional evidence for the human evaluation process by considering two individual-level AI-relevant characteristics (e.g., people's experience with LLMs and their perceived AI use of the work). Our work shows some level of human favoritism from the evaluators' side when evaluators are uninformed about the origins of the work they assess, which differs from existing work in which evaluators know whether the work they assess is AI-generated (Yin, Jia, and Wakslak, 2024).

1.2 Related Literature

1.2.1 LLMs' Generative Capabilities During Human-AI Collaboration

LLMs, built on transformer-based learning models trained on extremely vast internet-sourced datasets (i.e., Wikipedia and Google Images) (Minaee et al., 2024), represent a paradigm shift in artificial intelligence. LLMs operate by learning probability distributions over high-dimensional semantic spaces, enabling them to generate novel, contextually relevant outputs. Such generative capacity positions LLMs as unique tools for augmenting human work, particularly in creative and knowledge-intensive tasks where variability and adaptability are critical (Brynjolfsson, 2023).

Distinguishing model capabilities: Overall capabilities versus generative capabilities In this study, we first distinguish between *LLMs' overall capabilities* (differences in performance across model versions of LLMs, such as GPT-4 vs GPT-3) and *LLMs' generative capabilities* (the ability of a specific LLM, like GPT-3, to govern output diversity by adjusting parameters within the model). The former, LLMs' overall capabilities, has dominated researchers' attention with studies demonstrating the positive correlation between the size of a language model (i.e., the number of its parameters) and its performance (Brown et al., 2020). Allegedly comprising 1.76 trillion parameters—nearly ten times the 175 billion parameters of GPT-3.5—GPT-4 shows significantly superior performance across a range of academic and professional benchmarks, reflecting substantial gains in natural language processing capabilities attributable to its increased model scale (Brynjolfsson, 2023). However, limited studies investigate the influence of the latter term, *LLMs' inherent generative capabilities*, on shaping work performance.

Despite rapid advancements in LLMs, recent literature addresses a paradox in human-AI collaboration: improvements in the first term, *LLMs' overall capabilities*, do not consistently translate into better collaborative outcomes (Noy and Zhang, 2023; Lee et al., 2023; Chen and Chan, 2024; Li et al., 2024; Boussioux et al., 2024). For example, Lee et al. (2023) find that more advanced LLMs do not consistently enhance joint performance across tasks, while

Li et al. (2024) report only a slight gain when users collaborate with GPT-4 versus GPT-3.5. These findings suggest that technological improvements in *LLMs' overall capabilities* yield limited returns in human-AI collaboration contexts, indicating a complex interaction between the capabilities of AI models and human factors (Noy and Zhang, 2023). While prior work highlights the paradox of *LLMs' overall capabilities* within human-AI collaboration, it remains unclear whether similar findings exist when it comes to the second term, *LLMs' inherent generative capabilities.*

The underexplored frontier: LLMs' generative capabilities in the human-AI collaboration Existing literature has examined *LLMs' generative capabilities* in isolation, often manipulating parameters such as temperature to assess output diversity in tasks like narrative writing and design (Bellemare-Pepin et al., 2024; Peeperkorn et al., 2024). Research by Ma et al. (2024) explores these capabilities by comparing design solutions generated independently by AI and humans. These studies conceptualize AI as an independent agent and show that higher temperature values generally enhance creativity in AI-generated content. Yet, the impact of *LLMs' generative capabilities* within human-AI collaboration remains underexplored, despite this being the more prevalent context where the work is co-created through dynamic human–AI interaction.

Therefore, our study investigates how varying *LLMs' inherent generative capabilities* affect human-AI co-creation. Resolving this is important for theorizing how LLMs can be designed and deployed to complement, rather than conflict with, human creative processes. Specifically, we manipulate *LLMs' generative capabilities* using two adjustable parameters that can significantly affect the randomness of the text produced by LLMs[1]: temperature and

[1] https://platform.openai.com/docs/api-reference/audio

frequency penalty. On one hand, “temperature” in GPT acts as a control parameter for the randomness in generating text by flattening or sharpening the probability distribution over tokens. Higher temperature settings yield a flatter probability distribution, increasing lexical diversity and generating less predictable, more divergent text. In contrast, lower temperatures concentrate probability mass on high-likelihood tokens, resulting in more deterministic and predictable outputs. On the other hand, “frequency penalty” discourages repetition by reducing the likelihood of reusing previously generated tokens, thereby indirectly enhancing output diversity and promoting greater variation in the generated text. Moreover, differentiating from prior work using the objective assessment (Lee, Liang, and Yang, 2022), we propose a subjective assessment framework for creative narrative evaluation, which is crucial for creative tasks like narrative writing.

1.2.2 IT Affordance and Human Creators’ Algorithm Aversion During Human-AI Collaboration

Recent studies examine the impacts of integrating AI into human-centric workflows (e.g., Bauer, von Zahn, and Hinz (2023)). There have been experimental studies (e.g., Wang, Yang, and Sun (2023); Noy and Zhang (2023)) as well as field studies (e.g., Brynjolfsson (2023)), demonstrating that adopting LLMs into humans’ work can significantly enhance creative and operational performance. However, the IT affordance literature shows that such benefits are contingent on how users interact with IT artifacts (e.g., AI or LLMs) over time (Markus and Silver, 2008; Majchrzak and Markus, 2012). Improved performance through IT use arises not automatically but through users’ selective enactment of affordances. Individuals tend to exhibit algorithm aversion (Dietvorst, Simmons, and Massey, 2015, 2018), judging algorithms negatively a priori, thereby limiting their potential impact. Algorithm aversion is particularly common among experts, whose domain knowledge is often associated with reduced reliance on automated systems (Whitecotton, 1996; Commerford et al., 2022). This reluctance is partly driven by unrecognized overconfidence in intuitive judgment, a cognitive bias that diminishes perceived value in algorithmic support (Eining et al., 1997; Sieck and Arkes, 2005).

1.2.3 Human Evaluators' Characteristics During the Evaluation Process

Next, we examine the evaluation process by reviewing literature on two key characteristics of human evaluators: their experience with algorithms and the influence of human favoritism.

Experience with algorithms Prior literature differentiates between experience with the decision domain (experts in terms of domain knowledge) and specific experience with algorithmic decision aids (experts in terms of working with algorithms) (Burton, Stein, and Jensen, 2020), yet most research investigates the former, domain expertise. Existing literature shows that experts' tendency to prefer human judgment over algorithms in fields where individuals possess professional expertise (Montazemi, 1991; Whitecotton, 1996). Such bias is evident in fields like auditing and radiology, where professionals often disregard algorithmic evidence, particularly when it conflicts with their initial assessments and the AI process is opaque (Commerford et al., 2022; Lebovitz, Lifshitz-Assaf, and Levina, 2022). However, our understanding of how the latter term, experience with algorithmic decision aids, affects evaluative judgments remains limited.

Existing research on experience with algorithmic decision aids has explored how it influences trust and usage; however, much of this work has focused on users who directly collaborate with these systems, rather than those who evaluate their outputs. Specifically, studies suggest that human collaborators—individuals who interact directly with algorithmic decision aids—may develop greater trust and reliance on these systems as they become more familiar with them (e.g., Burton, Stein, and Jensen (2020)). However, the evidence is mixed and highly dependent on context (Turel and Kalhan, 2023). Notably, this body of work has paid relatively little attention to human evaluators, those tasked with judging the quality of algorithm-generated content. These evaluators may form their judgments through different cognitive and experiential processes. As LLMs become increasingly embedded in communication and decision-making, it is critical to understand how evaluators' prior experience with these systems shapes their subjective assessments, particularly in domains like creativity where human judgment plays a central role (Jakesch et al., 2023). Thus, while experience has been shown to affect how collaborators engage with algorithmic systems, we still know little about how it influences evaluators' judgments—an important gap given the growing presence of LLMs in shaping human opinion.

Human favoritism Recent literature suggests that human favoritism underlies biased evaluations against AI-generated content, particularly in subjective tasks (Castelo, Bos, and Lehmann, 2019; Morewedge, 2022). Individuals consistently rate creative outputs (i.e., paintings, poetry, and news articles) as less favorable when attributed to AI (Clerwall, 2017; Ragot, Martin, and Cojean, 2020; Köbis and Mossink, 2021). Similar findings exist for interpersonal communication, where AI-authored empathetic messages lose impact upon disclosure (Yin, Jia, and Wakslak, 2024), and for content quality assessments, which improve with human attribution (Zhang and Gosline, 2023; Millet et al., 2023).

However, there may be scenarios where humans do not disclose their use of AI, and the content created by a human and/or an AI becomes indistinguishable. When facing uncertainty about whether content is generated by humans, AI, or their collaboration, people often rely on quick, intuitive, and heuristic judgments to assess its quality (Hafenbrädl et al., 2016; Jarrahi, 2018). Our understanding of this phenomenon remains incomplete, especially in the case where AI assistance is undisclosed. Therefore, this study seeks to examine whether a bias favoring human-generated content persists even when the authorship source is unknown. Specifically, we investigate whether an increased perception of AI involvement potentially leads to a diminished perceived creativity of the work.

1.2.4 Narrative Creativity

Writing can be characterized as a creative process where the writer actively engages with the evolving text (Emig, 1971). This is because during this process, writers act as creative thinkers and problem-solvers who manage task constraints while utilizing the creative and linguistic resources available (Sharples, 2002; D'Souza, 2021). The literature has examined various adoption of LLM writing assistants in creative tasks, viewing LLMs not merely as tools for word prediction or correction but as active co-authors (Lee, Liang, and Yang, 2022; Yang et al., 2022; Yuan et al., 2022). Design characteristics for better interaction with writing assistants can support inspiration (Wang, Yang, and Sun, 2023; Bhat et al., 2023; Lee, Liang, and Yang, 2022), language proficiency (Buschek, Zürn, and Eiband, 2021), shorter and more predictable texts (Arnold, Chauncey, and Gajos, 2020), more standard phrase usage (Buschek, Zürn, and Eiband, 2021), or creative writing (Clark et al., 2018; Yuan et al., 2022). Bhat et al. (2023) examines how writers assess the suggestions provided and incorporate them into various cognitive writing processes.

Table 1.1 Related literature on human-AI collaboration and creativity.

Literature	**The role of LLMs** (LLMs' generative capabilities)	**The role of human collaborators** (human collaborators' AI use: -Binary: with AI vs without AI -Continuum: the extent to use AI)	**The role of varying human-AI collaboration** patterns (Dyadic collaboration modes such as mutuality and equality)	**Considering both human collaborators' characteristics (CC) and evaluators' characteristics (EC)**
Wang et al. (2023)	yes (AI-only, without human-AI collaboration)	binary	no	only CC
Bellemare-Pepin et al. (2024)	yes (AI-only, without human-AI collaboration)	binary	no	only CC
Lee et al. (2022)	yes (in the human-AI collaboration)	continuum (as a dependent variable)	yes (view each variable independently as a dependent variable)	only CC
Brynjolfsson et al. (2023)	no	binary	no	only CC
Jakesch et al. (2023)	no	binary	no	only CC
Noy and Zhang (2023)	no	binary	no	both
Zhang and Gosline (2023)	no	binary	no	both
Yin et al. (2024)	no	binary	no	only EC
Millet et al. (2023)	no	binary	no	only EC
Ragot et al. (2020)	no	binary	no	only EC
Köbis and Mossink (2021)	no	binary	no	only EC
Logg et al. (2019)	no	continuum	no	only CC
Turel and Kalhan (2023)	no	continuum	no	only CC
This study	yes (with the human-AI collaboration	continuum (as a moderator)	yes (define four types of human-AI collaboration modes based on these two variables	both

We summarize the empirical studies on human-AI collaboration in Table 1.1.

1.3 Hypotheses Development

We begin this section by outlining the research framework, followed by the development of hypotheses corresponding to each research question. Figure 1.1 summarizes our research framework, which serves as a roadmap for hypothesis development. Despite growing interest in human–AI co-creation, existing research on narrative creativity has not systematically examined how intrinsic characteristics of LLMs, such as their generative capabilities, interact with human creative processes to shape the perceived creativity of co-produced outputs. This gap is particularly crucial given that the so-called "black box" nature of LLMs may obscure important theoretical and practical implications for how AI augments human creativity. To address this, we investigate the relationship between inherent parameters of LLMs (specifically, randomness) and subjective evaluations of creative output. Our work thereby bridges computational specifications and creative judgments, offering empirical insight into how the generative properties of LLMs influence perceived creativity in co-created narratives.

In addition, we examine algorithm aversion in human-LLM collaboration by moving beyond the traditional expert/non-expert lens. Specifically, we investigate how individuals differentially utilize LLM-generated advice during creative tasks, where empirical evidence remains limited. While some collaborators avoid algorithmic input despite its potential benefits, others may over-rely on it, potentially stifling creativity. While prior literature has examined the effects of disclosed AI authorship (Zhang and Gosline, 2023) or human authorship (Gnewuch et al., 2024), many real-world settings involve ambiguity around AI involvement. When the origin of content is unknown, individuals often rely on heuristic judgments (Hafenbrädl et al., 2016; Jarrahi, 2018). Therefore, our work complements this stream of literature by examining whether perceived AI involvement lowers perceived creativity even when authorship is undisclosed. We investigate if a human-favoring bias persists under uncertainty, offering new insight into evaluative dynamics in AI-assisted creative contexts.

We first focus on the human-AI co-creation and examine whether LLMs' generative capability affects the perceived creativity of final co-created work (H1a), and how human creators' utilization of LLMs' suggestions moderates this relationship (H1b). Next, we study how human-AI collaboration patterns shape the perceived creativity of work (H1c), controlling for LLMs' generative capabilities and algorithm aversion among human creators. Finally, we turn to the human evaluation process and examine how human evaluators' characteristics (i.e., their familiarity with LLMs and perception of AI use) influence the perceived creativity of work (H2a, H2b). Hypotheses are developed in the following subsection.

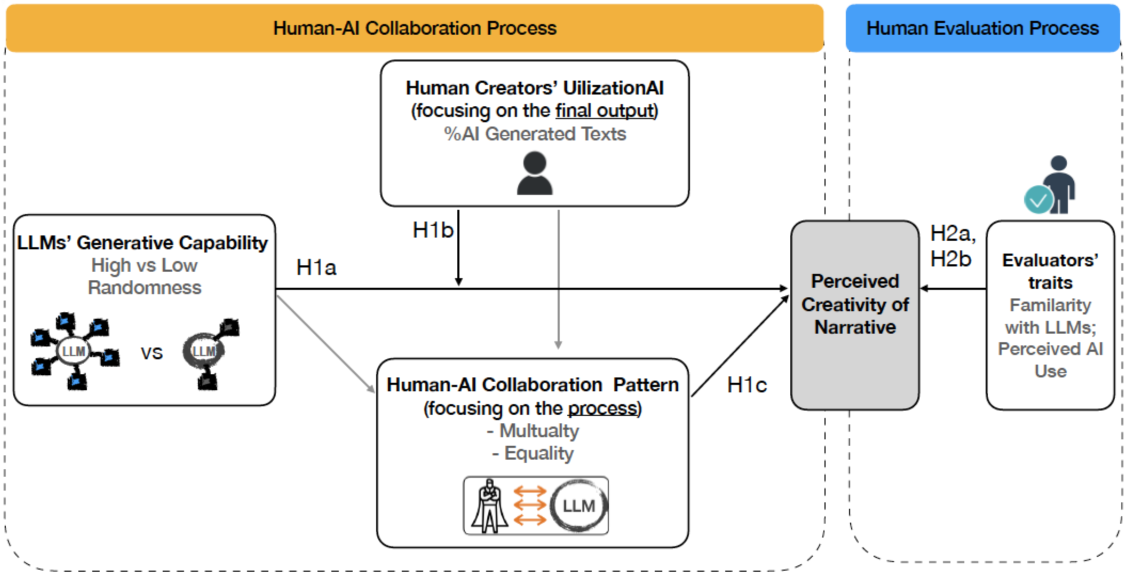


Figure 1.1 Research framework.

1.3.1 Within the Human-AI Collaboration Process

In creative writing, the generation of creative narratives involves not only organizing ideas into coherent plots but also writing stories with novelty and originality (McKee, 1997; Amabile, 2018). When LLMs are set to high randomness, they produce more diverse and unexpected suggestions (Roemmele and Gordon, 2018). Thus, when interacting with such suggestions, this greater diversity can augment human writers' divergent thinking—an essential component of creative writing (Runco and Acar, 2012)—by expanding their cognitive repertoire, stimulating curiosity, and encouraging the exploration of alternative approaches. In this context, human writers can integrate or synthesize these unique suggestions into their narratives, thereby enhancing the overall perceived creativity of the stories.

In contrast, when LLMs operate in a low randomness setting, their generated responses tend to be more predictable and conservative. Such outputs reinforce established patterns and limit the availability of novel ideas, which in turn encourages writers to adhere to familiar cognitive pathways rather than exploring unconventional perspectives. This reduction in cognitive variability has been shown to diminish the potential for creative ideation (Chakrabarty et al., 2024). As a result, the overall creativity of the narratives may diminish when human writers collaborate with low-randomness LLMs, compared to those generated in collaboration with high-randomness models. Accordingly, we hypothesize that in the context of human-AI co-creation, configuring LLMs with high randomness will serve as a catalyst for more innovative narrative construction, yielding stories that external evaluators judge as more creative.

H1a: *Collaborating with high randomness LLMs, which generate more diverse suggestions, leads to increased perceived creativity of work as assessed by external evaluators, compared to collaborating with low randomness LLMs.*

While high-randomness LLMs can offer a wide range of novel suggestions, human creators ultimately retain the right to which ideas to adopt or discard during the creation process. When it comes to tasks related to creative intelligence, individuals tend to exhibit algorithm aversion, stemming from concerns about personal identity and the longstanding belief that creativity is uniquely human (Morewedge, 2022). Thus, although collaborating with high-randomness LLMs exposes creators to a diverse range of unexpected suggestions, the enhanced perceived creativity of the final work depends on their actual integration of these unconventional ideas with their own insights.

Conversely, when collaborating with low-randomness LLMs, human creators tend to receive a narrower set of conventional suggestions that may discourage further creative exploration. According to cognitive miser theory (Orbell and Dawes, 1991), individuals prefer to solve problems using simpler and less effortful ways rather than more complex and demanding ones. As the ease of accessing assistance from LLMs increases, people who heavily rely on AI to complete tasks might choose to reduce their engagement and efforts in the human-AI collaboration process. As a result, in the context of creative writing, heavy reliance on the outputs of low-randomness LLMs can inadvertently restrict creative exploration and diminish the perceived creativity of the final output. Therefore, we hypothesize that the effect of LLMs' generative capabilities (i.e., high vs low randomness) on perceived creativity of work is moderated by the level of humans' utilization of LLMs' generated advice. Specifically,

H1b: *Greater utilization of advice generated by high (low) randomness LLMs, leads to increased (decreased) perceived creativity of work as assessed by external evaluators.*

Given that human creators tend to exhibit comparable levels of algorithm aversion, the nature of human–AI collaboration itself may play an important role in facilitating creative outcomes. Mutuality, defined as the iterative, active engagement between humans and AI, allows humans to dynamically interact with AI-generated suggestions. When human creators keep interacting with AI outputs, they can adapt and incorporate a diverse range of ideas, thereby establishing a productive feedback loop that enhances the overall creativity of work. Thus, a high level of mutuality (i.e., frequently seeking AI assistance and making deliberate selections) facilitates the merging of human intuition with AI's innovative suggestions, leading to greater creative work compared to a passive, one-off use of AI systems (Wan et al., 2024; Boussioux et al., 2024). In contrast, limited interaction between humans and AI restricts this iterative process, hindering the effective transformation of AI suggestions into innovative solutions and ultimately diminishing the perceived creativity of work. In creative writing tasks, for instance, human writers who actively collaborate with high-randomness LLMs are better positioned to explore a wide range of novel ideas and integrate these insights into their work. Thus, a high degree of mutuality in the human–AI collaboration tends to result in outputs that are perceived as more creative.

Furthermore, certain essential aspects of creativity production, such as the courage to face uncertainty and adversity, are uniquely human and beyond the capabilities of AI (May, 1994). The courage to create stems from existential struggles and personal experience, which AI may inherently lack. Although advanced language models like Claude and GPT-4 have demonstrated competence in divergent thinking and problem-solving, they still underperform in creative writing tasks (Sun et al., 2024). However, when humans take an active, leading role (i.e., revising AI-generated outputs), they can infuse the collaborative process with these uniquely human qualities. Therefore, on top of greater mutuality/interaction between humans and AI, such human-led collaboration not only leverages AI's generative potential but also integrates the creator's characteristics related to creativity production (i.e., the essential courage to create), enhancing the perceived creativity of the final output (Lockhart, 2024). Accordingly, assuming the same level of LLMs' generative capabilities and human creators' algorithm aversion, we hypothesize that:

H1c: *For a given level of LLMs' generative capabilities and human creators' algorithm aversion, the high-mutuality, high-human-led collaboration mode yields the highest perceived creativity compared to the low-mutuality, low-human-led collaboration.*

1.3.2 Within the Human Evaluation Process

Human evaluators who are more familiar with LLMs tend to have a personal trait of openness to experience with technologies, which enhances their appreciation for creative collaborative work. According to Rogers, Singhal, and Quinlan (2014), innovators and early adopters of new technology possess specific traits that make them more receptive to innovations. These traits include open-mindedness and a positive attitude towards change. In our context, people who are more familiar with a new technology (i.e., LLMs) tend to be more open-minded, implying they are curious and open to exploring new ideas and technologies. Open-mindedness can be considered the opposite of algorithm aversion. Open-minded individuals are more likely to evaluate written work based on its inherent qualities (i.e., perceived creativity), regardless of whether it was generated by LLMs. In contrast, those who are less open-minded may harbor prejudices against LLMs, often assigning lower ratings to content that is perceived to be mostly generated by LLMs. Therefore, individuals with more familiarity with LLMs tend to appreciate more the expressive qualities of narratives without concern for the nature of their authorship (human or LLM), leading to a greater appreciation of perceived creativity in the works they assess.

H2a: *People who are more familiar with LLMs tend to perceive greater creativity of work, compared to those who are less familiar with LLMs.*

When humans act as evaluators who assess the perceived creativity of written work, they are more likely to feel their identity is threatened in areas that are important to their personal identity, such as creative intelligence (Morewedge, 2022). Furthermore, Ornes (2019) argues that computers challenge the notion of creativity, once thought to be uniquely human. In our context, when it comes to evaluating narratives, creativity serves as a way to demonstrate human uniqueness and intelligence. Thus, despite the nature of the writer (human or LLM) being undisclosed, an increased suspicion of AI involvement in a creative narrative may intensify evaluators' feelings of threatened identity. This, in turn, could lead to a manifestation of algorithm aversion, which in our context is operationalized as diminished perceived creativity of the work.

H2b: *When evaluators suspect significant AI involvement in written work, despite the source being undisclosed, they tend to appreciate the work less, perceiving it as less creative.*

1.4 Data, Measures and Models Specification

1.4.1 Data

To answer these questions, we employ a quasi-experimental framework by focusing on creative writing tasks under two distinct conditions – low vs high randomness – during the human-AI co-creation. Specifically, we use the CoAuthor dataset (Lee, Liang, and Yang, 2022), which is designed to reveal GPT-3's generative capabilities for human-AI interactive writing. This dataset includes a creative writing task, which involves 10 creative writing prompts from the Writing-Prompts subreddit (see Online Appendix Table A1). The experimental design used by the authors to create the dataset provides an ideal setting to study the impact of AI randomness on creativity. Each writer was randomly assigned to a GPT exhibiting either low or high randomness, by varying two decoding parameters in the model: temperature and frequency penalty. Thus, for each story, we construct a variable, $HighRand$, to represent the randomness level of the GPT setting, which is 1 for high randomness conditions and 0 for conditions of low randomness.

Each session started with a prompt, and writers could freely write, request suggestions from GPT-3, accept or dismiss suggestions, and edit accepted suggestions (see Online Appendix Figure A1). Each writer could write a maximum of five stories per prompt. For every writing session, which corresponds to a specific combination of writer and prompt, a GPT configuration was randomly assigned. In other words, the same writer could be assigned to

Table 1.2 Overall statistics of creative writing sessions in the CoAuthor dataset.

Prompts	Writers	Sessions	Time (min)	Total Words (words)	Queries	Acceptance Rate (%)	Written by AI (%)
10	57	830	11.6	446	12.8	75.7	26.6

either a high or low LLM randomness setting. Table 1.2 shows summary statistics for the CoAuthor dataset. The dataset contains 830 writing sessions written by 57 writers from Amazon Mechanical Turk. On average, each writing session is 446 words long, contains 12.8 queries to the system, has an acceptance rate of GPT of 75.7% (how often writers accepted suggestions from GPT-3), and results in 26.6% of final text written by GPT (the proportion of the final text written by GPT-3 as opposed to human writers). In other words, on average, 73.4% of final texts were written by humans as opposed to GPT-3.

1.4.2 Measures

Measures used for the human-AI collaboration process Aligned with Turel and Kalhan (2023), we conceptualize algorithm aversion and appreciation on a continuum by measuring *AIUtilization*, the extent to which participants incorporate LLMs' responses into their outputs (Li et al., 2024). Specifically, for each narrative, we begin by identifying the shared words between the AI-generated responses and the participant's final creative output (i.e., the final version of the creative narrative). We then calculate the ratio of these shared words to the total word count of the participant's final creative narrative. The value of *AIUtilization* ranges from 0 to 1, with higher values indicating greater utilization of LLMs in the final co-created narratives.

Instead of focusing solely on the final narrative output, we capture humans' collaborative abilities based on two dimensions (Storch, 2002; Lee, Liang, and Yang, 2022), including

$Mutuality$ and $Equality$. The first dimension, $Mutuality$, represents the extent of a human participant's engagement with the AI agent (e.g., by clicking the help button to request AI assistance, navigating through suggestions provided by the AI, selecting a suggestion, or reopening the interaction screen to further engage with the AI). Following Lee, Liang, and Yang (2022), we measure $Mutuality$ based on the counts of two types of event blocks[2]: human-AI interaction event blocks ($EventBlock(HumanAI)$) and human-alone event blocks ($EventBlock(HumanOnly)$). Specifically, $EventBlock(HumanAI)$ refers to any of four behaviors, including navigating suggestions provided by AI, choosing any of AI's advice, reopening the AI helper, and inserting any texts. $EventBlock(HumanOnly)$ refers to any of three behaviors, including dismissing AI advice, deleting, or inserting any texts. Specifically, we then construct the variable $Mutuality$ as the percentage of human-AI interaction event block counts to the total count of both human-AI interaction event blocks and human-alone event blocks $Mutuality = \frac{\sum_i [e_i \in EventBlock(HumanAI)]}{\sum_i [e_i \in EventBlock(HumanAI)] + \sum_i [e_i \in EventBlock(HumanOnly)]}$. A $Mutuality$ value of 100 indicates the writer relied entirely on AI interaction, while a value of 0 means the writer did not interact with the AI at all.

The other dimension, $Equality$, represents the balance in contributions between human creator and AI agent in generating final outputs. Following Lee, Liang, and Yang (2022), we measure $Equality$ based on the counts of two types of event blocks: human efforts event blocks ($EventBlock(HumanEfforts)$) and AI efforts event blocks ($EventBlock(AIEfforts)$). Specifically, the $Equality$ is defined as inserting any texts by human writers, while

[2] Specifically, for each story-writing session that involves various events such as text insertions, deletions, and cursor movements, we categorize these into event blocks—deterministic, non-overlapping sequences of related events. For example, an event block labeled "choose" might include actions like "suggestion-select" followed by "suggestion-close".

$EventBlocks(AIEfforts)$ is defined as choosing any of AI advice. Thus, we construct a variable $Equality$ as $1 - |\frac{\sum_i[e_i \in EventBlock(HumanEfforts)] - \sum_i[e_i \in EventBlock(AIEfforts)]}{\sum_i[e_i \in EventBlock(HumanEfforts)] + \sum_i[e_i \in EventBlock(AIEfforts)]}|$. The $Equality$ value of 1 indicates perfect parity, where human and AI contributions are equal. Conversely, a value of 0 signifies that one party (human or AI) exclusively contributes to the output, while the other makes no contributions. Empirically, we observe that human efforts consistently exceed AI efforts across all writing sessions, as $\sum_i[e_i \in EventBlock(HumanEfforts)] > \sum_i[e_i \in EventBlocks(AIEfforts)]$ for all writing sessions. Thus, in our contexts, a smaller value of equality reflects human-led efforts, rather than AI-led efforts.

As illustrated in Figure 1.2, four distinct quadrants can be constructed based on the median value of these two variables, representing different modes of human-AI collaboration. We aim to explore which collaboration mode is most strongly associated with higher perceived creativity of the work. Detailed descriptions of each collaboration mode are provided in Section 1.4.3.

Measures used for the evaluation process To measure the perceived creativity of narratives, we recruited participants from Prolific, an online research platform where people can sign in to do surveys and experiments. At the beginning of the survey, participants are presented with a consent form, which specifies that they will receive compensation of $6 for participating in a study lasting approximately 30 minutes. Only participants who agree to the consent form can continue their surveys. An attention check question is included to ensure participants pay attention to the survey.

Each participant is asked to read five randomly selected stories out of 830. For each narrative, participants need to answer the same set of questions related to their perceptions

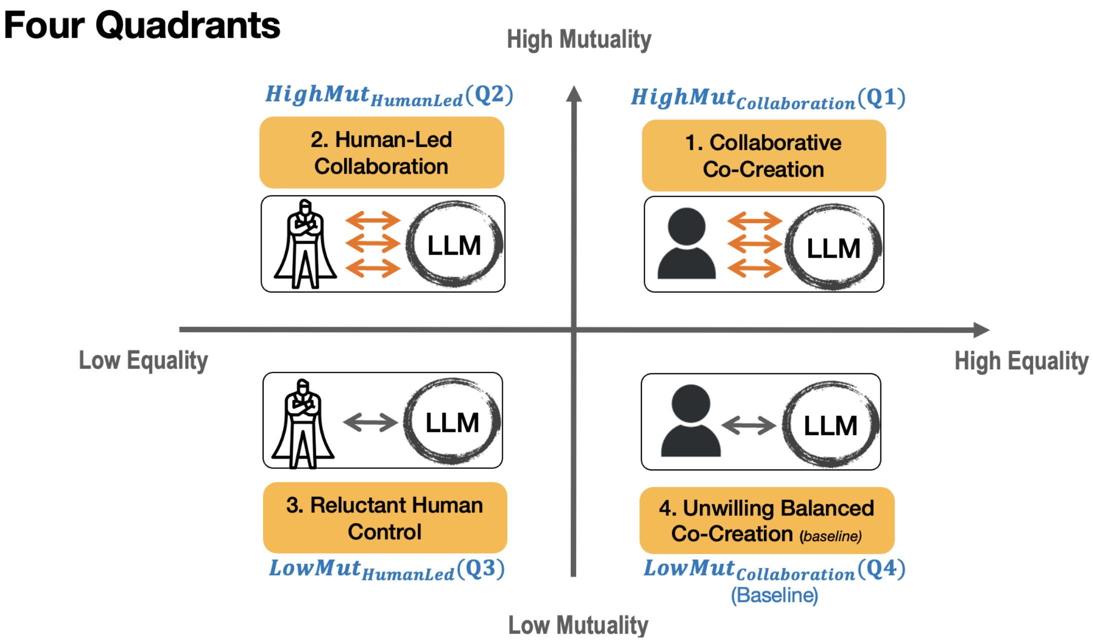

Four Quadrants
High Mutuality
$HighMut_{HumanLed}$(Q2)
2. Human-Led Collaboration
LLM
$HighMut_{Collaboration}$(Q1)
1. Collaborative Co-Creation
LLM
Low Equality
High Equality
LLM
3. Reluctant Human Control
$LowMut_{HumanLed}$(Q3)
LLM
4. Unwilling Balanced Co-Creation (baseline)
$LowMut_{Collaboration}$(Q4)
(Baseline)
Low Mutuality

Figure 1.2 Proposed four human-AI collaboration patterns based on Storch (2002).

of creativity and perceived AI use, as described in Table 3. Specifically, we construct a proxy of perceived creativity as follows. After reading each story, participants are required to answer four questions adapted from (Goncalo, Flynn, and Kim, 2010), which assess the level of creativity in the text using a scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). The four questions are outlined as follows: a) This article is creative; b) This article is more creative than other articles that have been recently published; c) Other people will think that this article is creative; d) It is unlikely that another author has come up with an article like this before. After examining whether the reliability of the scale and the inter-rater reliability is acceptable, the perceived creativity of each story is calculated based on the average ratings from the four questions and their respondents.

After labeling five stories, participants were asked about demographics such as age, highest level of education earned, employment status, and gender. Lastly, participants reported

Table 1.3 Variable definitions.

Variable	**Measurements**
Human as Creators: Human-AI Collaboration Process	
HighRand	High vs Low randomness GPT, by varying two decoding parameters in the model: temperature (T) and frequency penalty (FP). HighRand is 1 if the setting is (T=0.75; FP=1), and 0 if the setting is (T=0.3;FP=0).
AIUtilization	The ratio of the total number of LLMs' generated words over the total number of words for the story.
Mutuality	$Mutuality = \frac{\sum_i [e_i \in EventBlock(HumanAI)]}{\sum_i [e_i \in EventBlock(HumanAI)] + \sum_i [e_i \in EventBlock(HumanOnly)]}$, where *EventBlock(HumanAI)* refers to any of four behaviors, including navigating suggestions provided by AI, choosing any of AI advice, reopening the AI helper, and inserting any texts. *EventBlock(HumanOnly)* refers to any of three behaviors, including dismissing AI advice, deleting, or inserting any texts. A *Mutuality* value of 100 indicates the writer relied entirely on AI interaction, while a value of 0 means the writer did not interact with the AI at all.
Equality	$Equality = 1 - \vert \frac{\sum_i [e_i \in EventBlock(HumanEfforts)] - \sum_i [e_i \in EventBlock(AIEfforts)]}{\sum_i [e_i \in EventBlock(HumanEfforts)] + \sum_i [e_i \in EventBlock(AIEfforts)]} \vert$, where *EventBlock(HumanEfforts)* is defined as inserting any texts by human writers, while *EventBlock(AIEfforts)* is defined as choosing any of AI advice. An Equality value of 1 indicates perfect parity, where human and AI contributions are equal. Empirically, in our context, a value of 0 indicates that one party (human-led) exclusively contributes to the output, while the other (AI) makes no contribution. This is because in all writing sessions, total human efforts are greater than AI efforts.
Time	Total writing time (in min)
LogNumQuery	The logarithm of total number of queries from the LLM
Human as Evaluators: Evaluation Process	
Perceived Creativity	(Goncalo et al., 2010) Scale from 1 (Strongly Disagree) to 7 (Strongly Agree) a) This article is Creative; b) This article is more creative than other articles that have been recently published; c) Other people will think that this story is creative; d) It is unlikely that another author has come up with a story like this before.
Perceived AI Use	In your opinion, was the story generated by Artificial Intelligence (AI)? Responses are categorized as 1 for definitely AI-generated, 2 for maybe AI-generated, and 3 for definitely human-written. For ease of interpretation, we code *Perceived AI Use* as 3 if the rater chooses definitely AI-generated, 2 for maybe AI-generated, and 1 for definitely human-written. Thus, a higher value indicates a greater perception of AI use.
HF(FamiliarityLLMs)	On a scale from 1(Not at all familiar) to 5(Extremely familiar), indicate your level of familiarity with the uses of generative language model tools such as ChatGPT. We code *HF(FamiliarityLLMs)* as 1 if a rater's self-reported familiarity with LLM is 4 or 5, and 0 otherwise.
HF(FamiliarityAlg)	On a scale from 1(Not at all familiar) to 5(Extremely familiar), indicate your level of familiarity with algorithms and AI. We code *HF(FamiliarityAlg)* as 1 if a rater's self-reported familiarity with algorithms and AI is 4 or 5, and 0 otherwise.
Age	Assign a value of 1 if the evaluator is 45 years old or older, and 0 otherwise
Edu	Assign a value of 1 if the evaluator has earned a graduate degree (above a bachelor's degree) or higher, and 0 otherwise
FullTime	Assign a value of 1 if the evaluator is full-time employed, and 0 otherwise
Male	Assign a value of 1 if the evaluator is male, and 0 otherwise

their experience with LLMs by answering questions related to their familiarity with LLMs such as ChatGPT, and their familiarity with algorithms. We placed two questions regarding familiarity at the end of the survey to prevent them from influencing evaluators' perceptions of creativity during the initial labeling tasks. Table 1.3 summarizes definitions of the variables measured during the human-AI collaboration and evaluation processes.

1.4.3 Model Specification

To empirically investigate the impact of LLMs' generative capabilities on perceived creativity (H1a), we estimate the model in Equation (1.1). The dependent variable, Y_{ijp}, is the perceived creativity for story i generated by writer j based on prompt p. The independent variable $HighRand_{ijp}$ equals 1 if the GPT exhibits high randomness for story i with writer j and prompt p, and 0 otherwise. $Controls_{ijp}$ include two variables: total writing time to proxy the writer's effort ($Time$), and the logarithm of total number of queries to proxy the writer's willingness to interact with LLMs ($LogNumQuery$). We also control for writers' fixed effects λ_j to account for writer-invariant characteristics and the prompt's fixed effect α_p to account for time-invariant prompt characteristics that might affect perceived outcomes. Finally, we cluster standard errors at the writer level to address the problem of autocorrelation of error terms (Wooldridge, 2003).

$$Y_{ijp} = \beta_1 HighRand_{ijp} + \lambda_j + \alpha_p + Controls_{ijp} + \epsilon_{ijp} \tag{1.1}$$

To further examine the moderating effect of human creators' degree of algorithm aversion (H1b), we construct a variable, $AIUtilization_{ijp}$, that represents the ratio of the total number of LLMs' generated words over the total number of words for the final version of story i, which is written by writer j based on prompt p. We also add the interaction term $HighRand_{ijp} \times$

$AIUtilization_{ijp}$, as shown in Equation (1.2).

$$\begin{aligned} Y_{ijp} = \beta_1 HighRand_{ijp} + \beta_2 AIUtilization_{ijp} + \beta_3 HighRand_{ijp} \times AIUtilization_{ijp} \\ + \lambda_j + \alpha_p + Controls_{ijp} + \epsilon_{ijp} \end{aligned} \tag{1.2}$$

Moreover, to test H1c, we conduct a subsample analysis using Equation (3), applying it separately to two distinct settings: high-randomness and low-randomness. We explore which collaboration modes achieve the highest perceived creativity of work, given the same levels of LLMs' generative capabilities and human writers' algorithm aversion. For each setting (high-randomness or low-randomness), we define the dummy variable $HighAIUtilization_{ijp}$ based on the median value $AIUtilization_{ijp}$. This variable is assigned a value of 1 if $AIUtilization_{ijp}$ is greater than or equal to the median, and 0 if it is below the median. Moreover, for each setting, we define four distinctive collaborative modes (shown in Figure 1.2) based on the median values of $Multuality_{ijp}$ and $Equality_{ijp}$. Specifically, we construct four indicator variables, each representing a unique mode of collaboration: $HighMut_{Collaboration_{ijp}}(Q1)$, $HighMut_{HumanLed_{ijp}}(Q2)$, $LowMut_{HumanLed_{ijp}}(Q3)$, and Low $Mut_{Collaboration_{ijp}}(Q4)$. For example, a writing session characterized by high levels of both mutuality and equality would be classified under $HighMut_{Collaboration_{ijp}}(Q1)$, reflecting strong interaction and balanced participation between human creators and AI. As a result, the coefficient vector γ_1 in Equation (1.3) is 4-dimensional, with $LowMut_{Collaboration_{ijp}}(Q4)$, serving as the baseline category. We then introduce four interaction terms by multiplying $HighAIUtilization_{ijp}$ with each collaboration indicator, making the coefficient vector γ_2

in Equation (1.3) also 4-dimensional.

$$\begin{aligned}
Y_{ijp} = {} & \beta_2 HighAIUtilization_{ijp} \\
& + \gamma_1 \{HighMut_{Collaboration_{ijp}}, HighMut_{HumanLed_{ijp}}, LowMut_{HumanLed_{ijp}}, LowMut_{Collaboration_{ijp}}\} \\
& + \gamma_2 HighAIUtilization_{ijp} \times \{HighMut_{Collaboration_{ijp}}, HighMut_{HumanLed_{ijp}}, LowMut_{HumanLed_{ijp}}, \\
& LowMut_{Collaboration_{ijp}}\} + \lambda_j + \alpha_p + Controls_{ijp} + \epsilon_{ijp} \qquad (1.3)
\end{aligned}$$

Lastly, we analyze the human evaluation process by estimating Equation (4) at the evaluator-story level. To test H2a, we investigate whether evaluators with lower familiarity with LLMs perceive human–AI collaborations as less creative. The dependent variable Y_{ir} is perceived creativity for story i rated by evaluator r. We include $HF(FamilarityLLMs)_r$, an indicator equal to 1 if the evaluator r reports high familiarity with LLMs (Likert scale 4 or 5), and 0 otherwise. The corresponding coefficient γ_F captures its average effect on the perceived creativity of work. To test the impact of evaluators' perceived AI use (H2b), we construct $PerceivedAIUse_{ir}$, representing rater r's perception of AI use when evaluating story i, ranging from 1 to 3, with larger values indicating greater perceived AI use. The coefficient of interest is γ_A reflects its effect on their perceived creativity of work. We also control for the story fixed effect τ_i to account for time-invariant story characteristics that might affect perceived outcomes[3]. $Controls_r$ represents a set of evaluator r's demographics, including age, gender, education, and employment status.

$$Y_{ir} = \gamma_F HF(FamilarityLLM)_r + \gamma_A PerceivedAIUse_{ir} + \tau_i + Controls_r + \epsilon_{ipr} \qquad (1.4)$$

[3] Because we already control the story fixed effect τ_i, we exclude inherent story-related variables such as $HighRand$, $AIUtilization$, and their interaction terms in the equation.

1.5 Empirical Results

1.5.1 Descriptive Statistics

After conducting the survey on Prolific, 29 surveys were disqualified for failing attention checks, 4 were rejected for being completed too quickly, and 4 timed out, ending up with 479 survey submissions. Because each participant was assigned five randomly selected stories, it was possible that a small set of stories would not be read. In our case, out of 830 stories, one was read only once, while the remaining 829 stories were read a minimum of two times and a maximum of ten times. Thus, we select those 829 stories as the sample for the following analysis.

Before computing the aggregated perceived creativity score for each story, we first determined consistency among raters. Specifically, we obtained an average inter-rater reliability (IRR) of 0.68, which is considered "moderate agreement" (Landis and Koch, 1977; Siegert, Böck, and Wendemuth, 2014). We also computed Cronbach's alpha to assess the correlation among the four questions composing perceived creativity scale for each story. In our sample, Cronbach's alpha was 0.87, implying that the four questions reliably measure the same construct. Therefore, for each story, we computed the average for each of the four items among raters, and then the average among these four items, utilizing the resulting metric as the unidimensional perceived creativity measure for each story.

Table 1.4 and Table 1.5 show the descriptive statistics of variables for the story-level analysis in the writing process (H1a-c), and for the evaluator-story-level analysis during the evaluation process (H2a-b), respectively.

Table 1.4 Summary statistics of variables in the writing process at the story level.

Variable	N	Mean	Std. dev.	Min	Max
Measures During Human Evaluation Process(Average)					
Perceived Creativity	829	4.44	0.92	1.5	7
Measures During Human-AI Collaboration Process					
AIUtilization	829	0.27	0.16	0	0.93
HighRand	829	0.49	0.50	0	1
Mutuality	829	0.38	0.03	0.33	0.48
Equality	829	0.28	0.16	0	0.80
Time	829	11.62	2.68	5.32	33.33
LogNumQuery	829	2.43	0.65	0	3.99

Table 1.5 Summary statistics of variables in the evaluation process at the rater-story level.

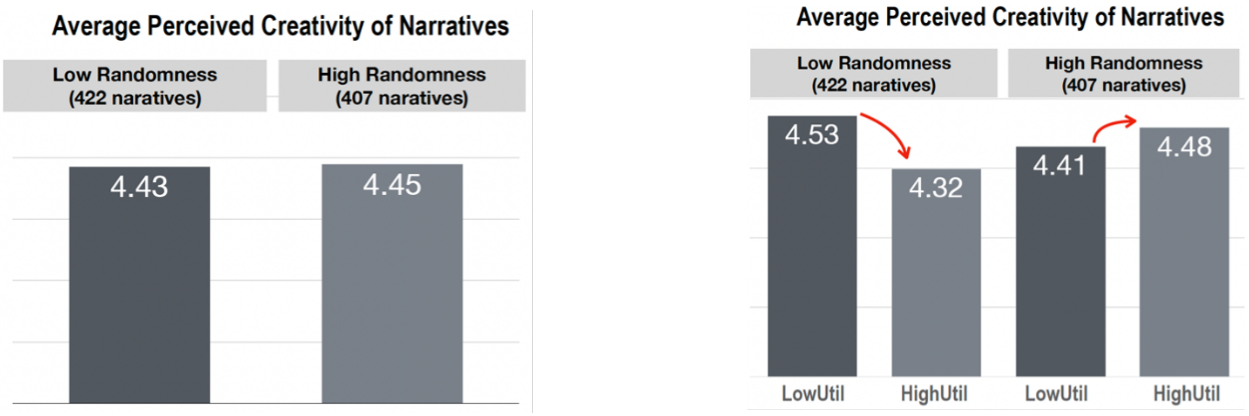
Variable	N	Mean	Std. dev	Min	Max
Perceived Creativity	2,389	4.43	1.36	1	7
Perceived AI Use	2,389	1.95	0.71	1	3
HF(FamiliarityLLMs)	2,389	0.40	0.49	0	1
Age	2,389	0.38	0.49	0	1
Edu	2,389	0.53	0.50	0	1
Fulltime	2,389	0.56	0.50	0	1
Male	2,389	0.53	0.50	0	1

1.5.2 Hypotheses Tests

Within the Human-AI Collaboration Process (H1a-c Tests)

We first provide model-free evidence for H1a-b. Figure (1.3a) shows the average perceived creativity of narratives produced under varying levels of GPT randomness during human-AI collaboration. The results reveal negligible differences in average creativity ratings between low randomness (4.43) and high randomness (4.45) conditions. This minimal difference ($\Delta = 0.02$) suggests that the increased LLM randomness does not directly enhance the perceived creativity of co-generated outputs. Such findings indicate the failure to support H1a, which posited that collaborating with higher LLM randomness would yield more creative narratives.

Moreover, for easy visualization, we categorize the sample narratives into two groups based on the median value (0.23) of the continuous variable AIUtilization: high and low



(a) (b)

Figure 1.3 Model-free evidence of H1a-b. Average perceived creativity of narratives for different levels of GPT randomness (left, Figure 1.3(a)) and writers' AI utilization (right, Figure 1.3(b)) during the human-AI collaboration.

degree of AI utilization. Figure (1.3b) shows average perceived creativity for different levels of GPT randomness and writers' utilization of AI's advice during the human-AI co-creation. Under low GPT randomness, narratives that heavily utilize LLMs' advice (with the average value of 4.32) are perceived as less creative than those with minimal LLM utilization (with the average value of 4.53). Conversely, under high GPT randomness, narratives with high AI utilization (with the average value of 4.48) are perceived as slightly more creative than those with low AI utilization (with the average value of 4.41). Thus, this trend indicates that the level of human creators' AI utilization shapes the impact of GPT randomness during human-AI co-creation.

Column (1) of Table 1.6 presents the estimation results from (1.1), where the primary independent variable is $HighRand$. The coefficient of $HighRand$ in Column (1) is positive but insignificant, indicating **a lack of evidence to support H1a**. During the human-AI collaboration, increasing an LLM's randomness is not sufficient in generating more creative narratives. Therefore, the lack of a statistically meaningful effect calls into question assumptions about the influence of LLMs' inherent generative capabilities (e.g., randomness) on creative outcomes in human-AI collaboration, highlighting the need to reconsider how these model attributes interact with human creative processes.

Then, we investigate the moderating role of human creators' algorithm aversion using Equation (1.2), which adds $AIUtilization$ and the interaction term $HighRand \times AIUtilization$. The results are shown in Column (2) of Table 1.6. The estimated coefficient for $HighRand$ (-0.023) is negative and statistically significant. When human creators show full algorithm aversion—defined as a rejection of AI-generated inputs in creative decision-making (i.e., $AIUtilization = 0$)—into the creative process, collaborating with high-randomness AI may hinder the user's ability to craft creative narratives. Potential mechanisms could be high-randomness AI, which prioritizes divergent, unpredictable outputs, amplifying cognitive friction for human creators already predisposed to dismissing algorithmic contributions. Such diverse advice provided by AI further imposes a *high technology overload* on creators (Bunjak, Černe, and Popovič, 2021) who must reconcile dissonance between their intent and the AI's incongruent suggestions, thereby leading to burnout and lower perceived creativity of work.

Importantly, the coefficient of the interaction term $HighRand \times AIUtilization$ (0.965) is positive in Column (2) with full controls, suggesting a significant moderating effect of human creators' algorithm aversion during human-AI collaboration. When the LLM generates more diverse suggestions (in the high-randomness GPT setting), stories generated with greater AI utilization are perceived as more creative, compared to those narratives with less AI utilization. Taken together, we find that the impact of LLMs' generative capabilities on the perceived creativity of work varies among human writers' utilization of LLMs' suggestions,

Table 1.6 Effects of LLM's generative capabilities and human creators' AI utilization on perceived creativity of work (testing H1a and H1b).

	Dependent variable: Perceived Creativity	
	Effect of LLMs' Randomness	Effect of LLMs' Randomness and AI Utilization
	(1)	(2)
$HighRand$	0.016	-0.223**
	(0.060)	(0.107)
$AIUtilization$		-1.026*
		(0.575)
$HighRand \times AIUtilization$		0.965***
		(0.333)
Controls	Yes	Yes
Prompt FE, Writer FE	Yes	Yes
Observations	821	821
R^2	0.093	0.101

Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Controls include writers' total writing time, the logarithm of total number of requests from LLMs, prompt-fixed effect, and writer-fixed effects. Cluster-robust standard errors at the individual writer level are shown in parentheses.

which **supports H1b**.

We further investigate which human-AI collaboration modes can trigger greater perceived creativity of work (H1c). Table 1.7 represents the estimation results of Equation (1.3). In both subsamples, high randomness LLMs (Column 1) and low randomness LLMs (Column 2), the interaction terms of $HighAIUtilizationXHighMut_{HumanLed}$ are both positive and statistically significant. Therefore, for a given level of LLMs' generative capabilities and relatively high levels of human creators' AI utilization, the high-mutuality, high-human-led collaboration mode yields the highest perceived creativity compared to the low-mutuality, low-human-led collaboration, which **supports H1c**.

In addition, compared to the baseline interaction mode ($LowMut_{Collaboration}$(Q4)), the second most effective human-AI interaction pattern is $HighMut_{Collaboration}$. This is evidenced by the positive and statistically significant coefficients of the interaction term $HighAIUtilization \times HighMut_{Collaboration}$, which are 0.969 in the high-randomness setting (Column 1) and 0.905 in the low-randomness LLM subsample (Column 2). These findings suggest that greater mutuality in human-AI interactions—whether through $HighMut_{HumanLed}$ or $HighMut_{Collaboration}$ mode—enhances the perceived creativity of the work.

Table 1.7 Effects of varying human-AI collaboration modes on perceived creativity of work (testing H1c).

	Dependent variable: Perceived Creativity	
	Subsample: High Randomness (1)	Subsample: Low Randomness (2)
$HighAIUtilization$	-0.426	-1.026**
	(0.268)	(0.575)
$HighAIUtilization \times HighMut_{Collaboration}$	0.969***	0.905**
	(0.331)	(0.430)
$HighAIUtilization \times HighMut_{HumanLed}$	1.316***	1.164***
	(0.415)	(0.376)
$HighAIUtilization \times LowMut_{HumanLed}$	0.629*	0.725
	(0.363)	(0.333)
Controls	Yes	Yes
Prompt FE, Writer FE	Yes	Yes
Observations	396	415
R^2	0.141	0.211

Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Controls include writers' total writing time, the logarithm of total number of requests from LLMs, prompt-fixed effect, and writer-fixed effects. Cluster-robust standard errors at the individual writer level are shown in parentheses.

Within the Human Evaluation Process (H2a-b Tests)

So far, our findings primarily stem from the collaborative work generation process involving humans and LLMs. In the following subsection, we investigate the evaluation process of narratives by focusing on external human evaluators. Specifically, whether humans exhibit

Table 1.8 Model-free evidence for H2a-H2b. Average perceived creativity among evaluators with high/low familiarity with LLMs and high/low perceived AI use.

	HF(FamilarityLLMs)		*Perceived AI Use*	
	Low	High	Low	High
Mean Perceived Creativity	4.36	4.53	4.57	3.94

Note: $HF(FamilarityLLM)$ is 1 if raters' familiarity with LLMs is 4 or 5, and 0 otherwise. For easy visualization, $PerceivedAIUse$ is considered high when it equals 3, and low when it equals 1 or 2.

a diminished appreciation for human-AI collaborative work when they are less familiar with LLMs (H2a) or when they believe the work is more likely generated by AI (H2b). Unlike previous analyses that are at the story level, the following analysis is conducted at the evaluator-story level.

Table 1.8 presents model-free evidence for H2a-H2b, which represents the average perceived outcomes among evaluators, categorized by varying levels of familiarity with LLMs and their perceived AI use. Evaluators who are highly familiar with LLMs assign a higher average creativity rating of 4.53, in contrast to those less familiar, who average a rating of 4.36. Both results of this model-free evidence support H2a. Moreover, narratives perceived to have low AI use generally receive higher ratings for perceived creativity than those with high perceived AI use, supporting H2b.

Model estimation results are shown in Table 1.9. The coefficients of $HF(FamilarityLLM)$ are positive and statistically significant in Columns (1-2), without and with the story-fixed effect, respectively. These findings demonstrate that evaluators who are more familiar with LLMs tend to perceive greater creativity in the evaluated work, compared to those who are less familiar with LLMs, which **supports H2a**. Furthermore, Columns (3-4) present negative and significant coefficients of $PerceivedAIUse$, both without and with the story-fixed effect (with values of -0.330 and -0.547, respectively). The results indicate that when evaluators suspect significant AI involvement in the creation of written work, despite the source being undisclosed, they tend to appreciate the work less, perceiving it as less creative, thus **supporting H2b**. Our findings complement existing literature on human favoritism (Yin, Jia, and Wakslak, 2024) by providing evidence that human biases can affect judgment, especially when the origin of creators is not disclosed. Lastly, the results remain consistent in Column (5) when considering both characteristics simultaneously. Thus, our study provides evidence that individuals' interaction with LLMs could also affect their decision-making process, particularly in evaluative judgments related to creativity assessment.

Table 1.9 Effects of evaluators' familiarity with LLMs and perceived AI use on perceived creativity of work as assessed by external evaluators. The analysis is at the evaluator-story level (testing H2a and H2b).

	Dependent variable: Perceived Creativity				
	Effect of Evaluators' Familarity with LLM		Effect of Evaluators' Perceived AI Use		Effect of Evaluators' Both Factors
	(1)	(2)	(3)	(4)	(5)
HF(FamiliarityLLM)	0.218*** (0.058)	0.144** (0.068)			0.158** (0.067)
Perceived AI Use			-0.357*** (0.038)	-0.327*** (0.046)	-0.330*** (0.046)
Age	0.300*** (0.058)	0.348*** (0.069)	0.256*** (0.057)	0.303*** (0.068)	0.323*** (0.068)
FullTime	0.219*** (0.058)	0.215*** (0.069)	0.229*** (0.058)	0.213*** (0.068)	0.210*** (0.068)
Edu(*AboveBachelor*)	-0.163*** (0.057)	-0.148** (0.066)	-0.154*** (0.056)	-0.132** (0.065)	-0.139** (0.065)
Male	-0.090 (0.057)	-0.002 (0.068)	-0.066 (0.055)	0.013 (0.066)	-0.004 (0.066)
Story FE	No	Yes	No	Yes	Yes
Observations	2389	2389	2389	2389	2389
R^2	0.020	0.430	0.049	0.446	0.448

Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regarding other demographic traits, the significant positive coefficients of *Age* and *Full Time* (Columns 1–5) suggest that individuals aged 45 and above, as well as those employed full-time, are more likely to report higher perceived creativity of work. Conversely, the significant negative coefficient for *Edu* suggests that higher levels of education are associated with lower perceived creativity. This result is consistent with Benedek et al. (2021), which found that individuals with more years of education are less likely to exhibit lower agreement on creativity myths, implying a lower probability of perceiving creativity in general. Moreover, we do not observe heterogeneous effects across gender, indicating that perceived creativity does not differ systematically between male and female evaluators.

1.6 Discussion and Implications

Building on prior research on human-in-the-loop in decision-making (Zanzotto, 2019; Ge et al., 2021; Fügener et al., 2022; Bauer, von Zahn, and Hinz, 2023), our findings show that the effect of LLMs' inherent generative capabilities on the perceived work creativity is crucially moderated by the human creators' utilization of AI. Specifically, higher randomness LLM enhances perceived creativity of output only when the human writer exhibits lower algorithmic aversion and is thus more receptive to the LLM's suggestions. In addition, our findings further address that effective human–AI collaboration in creative tasks hinges on active human engagement, highlighting the importance of interaction beyond the capabilities of the model itself. Through an analysis of varying collaboration patterns, we show that both the intensity of human–AI interaction and the extent of human-led contribution significantly affect the perceived creativity of the final output. Specifically, when human writers more frequently engage with LLMs (e.g., by asking follow-up questions and refining responses) and contribute to the tasks (e.g., by inserting more text), the resulting work is viewed as more creative. These results suggest that creativity is enhanced not only through iterative engagement with AI, but also when humans maintain a leading role in the creative process. Collectively, our study points to the importance of designing AI-assisted workflows that encourage sustained and dynamic human–AI collaboration to maximize the creative potential of LLMs.

Furthermore, as LLMs become more involved in everyday activities, understanding how individual-level characteristics related to LLMs shape human judgment becomes important. In this study, we examine how two specific evaluator attributes, their familiarity with AI and their perception of AI use in a given task, affect assessments of creativity in narrative writing. Our analysis yields two main insights. First, individuals with greater experience using LLMs are more likely to perceive higher levels of creativity in the texts they evaluate, suggesting that familiarity with AI may enhance sensitivity to creative nuances. Second, even when the authorship source is undisclosed, evaluators' assumptions about whether AI was involved significantly shape their judgments of creativity. This finding extends existing research on human favoritism by examining contexts where AI participation is ambiguous or inferred rather than explicitly stated. Together, these results address a critical need for caution: human evaluations of creative work may be systematically shaped by both prior exposure to AI and subjective assumptions about its authorship.

1.6.1 Implications for Research

Our research makes several important contributions to the growing literature on human–AI collaboration in creativity domains. While prior studies have examined the role of humans in shaping creative outcomes through prompting strategies (Wu, Terry, and Cai, 2022) and the functional roles individuals play in co-creative (i.e., ghostwriters who contribute content and sounding boards who provide evaluative feedback) (Chen and Chan, 2024), less attention has been paid to the inherent generative capabilities of LLMs themselves. LLMs are not passive tools; they possess generative properties that meaningfully influence creative outcomes. Recognizing LLMs as active agents in the creative process is essential to understanding the dynamics of co-creation. Our study is the first to empirically investigate how intrinsic characteristics of LLMs (particularly the randomness) affect the perceived creativity of co-produced content. These findings suggest the need to consider both human strategies and model-driven variability in interpreting divergent collaborative creativity.

In addition, we extend the literature on algorithm aversion in human–AI collaboration by moving beyond its traditional emphases on algorithm aversion (Logg, Minson, and Moore, 2019; Fügener et al., 2022; Turel and Kalhan, 2023) to investigate its moderating role in shaping perceived creativity of co-created work. Rather than conceptualizing algorithm aversion as a binary trait, we reconceptualize it as a continuous measure—captured by the extent to which human creators integrate AI-generated suggestions into their final work (Turel and Kalhan, 2023). Our findings reveal a nuanced interaction between AI system design and user behavior: when GPT operates with high randomness, greater AI utilization leads to increased perceived creativity. In contrast, under low randomness (i.e., more deterministic response), increased AI utilization corresponds with lower creativity evaluations. These results suggest that aligning the degree of human openness to algorithmic output with LLMs' internal generative properties is critical for optimizing co-creative outcomes.

Moreover, while prior work has categorized modes of human–AI collaboration (Revilla et al., 2023), our research extends theories of human–human co-creation (Storch, 2002) to the human–AI domain. We focus specifically on the dimensions of *mutuality* and *equality*: high mutuality ensures iterative alignment between human inputs and AI outputs, while maintain-ing a human-led collaboration mode safeguards contextual appropriateness and maximizes the creativity of work.

We further contribute to the literature on human-AI collaboration by introducing a dual-perspective framework for studying LLMs, shifting analytical attention from the process of AI-assisted creation to the human evaluation of AI-involved content. Specifically, we empirically examine how two evaluator-level factors—familiarity with LLMs and perceived AI involvement (in the absence of explicit disclosure)—shape creativity assessments. Extending prior work on AI's impact on human judgment (Jakesch, Hancock, and Naaman, 2023) and human favoritism (Logg, Minson, and Moore, 2019; Morewedge, 2022), our findings reveal that higher LLM familiarity and lower perceived AI use are both associated with increased perceived creativity of work. These results address the presence of evaluative biases in human–AI co-creation and highlight the critical role of individual experience and perception in shaping judgments of creative work.

1.6.2 Implications for Practice

Our research findings have important practical implications for policymakers and managers designing and implementing human-AI collaboration systems (Anthony, Bechky, and Fayard, 2023), particularly in tasks involving creative production and subjective evaluation. For policymakers and managers engaged in creative production tasks, our findings challenge a widely held assumption about LLMs: the greater randomness in their outputs inherently leads to more creative results. While it is commonly believed that increasing randomness fosters divergent thinking and enhances creative performance, our empirical evidence suggests otherwise. We find that higher randomness does not consistently translate into higher perceived creativity. Therefore, managers should be cautious about using randomness as a default lever for boosting creativity in AI-assisted work. Instead, we promote the development of structured frameworks that strategically adjust LLM parameters to align with the specific requirements of a given creative task. Thoughtful calibration, rather than blind reliance on randomness, is more likely to yield meaningful and effective creative outcomes.

Additionally, we encourage policymakers and managers to reconceptualize AI as a collaborative partner rather than a passive tool (Anthony, Bechky, and Fayard, 2023), drawing on the human-human collaboration literature to inform human-AI interaction design. By incorporating relational constructs such as mutuality and equality, our study extends prior work and reveals that these dimensions critically shape co-creative processes, often overlooked in outcome-focused studies (Wu et al., 2021; Sun et al., 2024). Our findings suggest that organizations should design AI systems that promote iterative interaction (e.g., click-to-refine prompts) and train employees to engage with AI as a "creative sparring partner." This relational framing offers a foundation for future research across modalities and cultural contexts, where collaboration norms may diverge.

For policymakers and managers involved in evaluating creative work, our research suggests the important role of human perception in assessing AI-assisted outputs. Specifically, we find that an evaluator's familiarity with LLMs significantly shapes their creativity judgments. Individuals with more experience using LLMs are generally more receptive to AI-generated contributions, rating them more favorably. In contrast, when evaluators merely perceive that AI was involved, regardless of whether it actually was, they often rate the output as less creative. These findings highlight the influence of cognitive biases and individual traits, such as a preference for human effort or skepticism toward AI, on subjective evaluation. As LLMs become more deeply integrated into routine workflows, yet humans remain central to interpretation and judgment, it is essential for organizations to recognize and actively manage these biases. Doing so is key to ensuring fair, consistent, and credible evaluations of AI-generated content.

1.6.3 Limitations

While this study provides valuable insights into the effect of LLMs' inherent generative capabilities in creative domains, it is important to acknowledge several limitations. Our analysis focuses exclusively on a single model (i.e., OpenAI's GPT-3) and a specific creative writing task. Future research should examine more recent LLMs to assess the generalizability of these findings. Moreover, the study employs fixed initial prompts and does not explore how prompt engineering might interact with model behavior. Subsequent work could investigate how variations in prompt design (e.g., Boussioux et al. (2024)), combined with different levels of model randomness, influence performance in human-AI collaborative settings. Finally, the optimal degree of randomness may vary across the domain and task context. For example, industries such as financial services may benefit from low randomness to maintain precision and reliability, whereas sectors like advertising or entertainment may derive value from higher randomness to enhance creativity and user engagement. Future empirical research across a broader range of domains is necessary to better understand these dynamics.

1.7 Conclusion and Future Research

It has been suggested that processes that were considered uniquely human, such as creativity and intuition, are being increasingly augmented by the speed, scalability, and analytical power of AI (e.g., Vaccaro, Almaatouq, and Malone (2024)). Given that the contours of creative tasks are significantly broadened with widespread deployment of AI models, what frameworks are by which we understand AI-human collaboration in creative tasks? Our findings provide insights into the process of human-AI collaboration by examining the role of LLMs' inherent generative capabilities and the mutuality of human-AI interaction in creative task performance. While these inherent capabilities enable more efficient deployment of AI assistance in creativity-driven workflows, our findings indicate that such parametric attributes alone (e.g., adjusting LLMs' inherent generative capabilities) may be insufficient to achieve optimal creative outcomes within human-AI collaboration. With increasing evidence that human-AI collaboration is key to the future of work, our study considers a dual perspective of mutuality and equality and extends prior theories of human collaboration to that of humans collaborating with AI.

Several future research directions could be explored based on our findings. Firstly, future work could examine the generalizability of our findings within human-AI interaction patterns across other forms of creative tasks, such as marketing ideation, artistic co-creation, and product innovation. Such work would help delineate the boundary conditions of collaborative creativity involving LLMs. Additionally, future research could explore the role of LLMs' generated capabilities and varying degrees of mutuality in human-AI collaboration in business contexts that, while not fully creative, still rely heavily on human judgment. For instance, tasks such as inventory ordering by human managers may be influenced by the nature of AI recommendations (Lu et al., 2025). Understanding how different configurations of AI-human interdependence affect decision quality in such contexts remains an open and important question. Another valuable direction for research could focus on the heterogeneous effects of LLMs on human collaborators. Research could investigate how LLMs influence the perceived creativity of outcomes, considering other psychological aspects of human collaborators such as irrationality (e.g., Shen, Jiang, and Zheng (2025)). Lastly, we consider the temperature parameter in this study, but empirical research could assess how variations in model-level parameters (e.g., fine-tuning, alignment processes) shape collaborative dynamics and outcomes in human-AI interactions. Given the evolving capabilities and deployment contexts of generative AI technologies, such research aligns well with the practice of responsible AI (Susarla et al., 2023).

CHAPTER 2 *GPT-DATector*: Increasing Accuracy And Decreasing Bias In GPT Detectors Using Creativity Measures

2.1 Introduction

Generative Artificial Intelligence (GenAI) tools such as ChatGPT have shown tremendous promise in supporting human endeavors across a wide range of domains, from e-commerce (Ghaffari, Yousefimehr, and Ghatee, 2024) and healthcare (Sharma et al., 2023) to education (Koltovskaia, 2020; Bubeck et al., 2023; Zhang et al., 2023; Yang et al., 2023). Such human-AI collaboration could potentially enrich the quality and creativity of work beyond what is produced by humans alone (Brynjolfsson, Li, and Raymond, 2025; Zhou and Lee, 2024). For instance, human-AI collaboration outperforms human or AI performance alone in image classification when the AI delegates tasks to humans; however, human delegation to AI is less effective due to humans' limited metaknowledge, or inability to accurately assess their own capabilities, leading to less optimal delegation (Fügener et al., 2022). However, when not effectively implemented or monitored, human-AI collaboration may turn into total reliance of humans on GenAI, which is especially concerning in fields such as education, where the proper development and nurturing of critical skills of the younger generation is at stake.

To address this challenge, one approach is to rethink conventional forms of educational assessment (Yeadon et al., 2023), advocating for changes to traditional assignments rather than relying solely on writing tasks, something GenAI can independently excel at. A complementary approach involves equipping educators with the tools to distinguish between human-written and GenAI-generated content. Given the limitations of human judgment in detecting AI-generated content (Jakesch, Hancock, and Naaman, 2023), as well as the biases in existing automated detectors (Liang et al., 2023), it is essential to develop effective automated GenAI detection tools across a wide range of educational writing tasks to mitigate potential academic misconduct and to ensure students develop fundamental critical thinking and writing skills (Susnjak and McIntosh, 2024).

Here, we focus on the latter approach, which maintains conventional text-based school assignments, such as argumentative essays, and aims at detecting the use of GenAI. Industry has recently created and open-sourced technology to watermark AI-generated content (Dathathri et al., 2024) but this technology does nothing to solve the problem of identifying all the content that was generated by GenAI before its availability – and until regulatory frameworks are defined and implemented, the level of adoption remains unclear. While there is growing interest in academia and industry to advance research in detecting text generated by GenAI (Tian and Cui, 2023; Venkatraman, Uchendu, and Lee, 2024), current detection tools are flawed, and the evolving developments of GenAI present a continuous challenge to effectively differentiate between human- and AI-generated content (Sadasivan et al., 2023). For instance, OpenAI launched a classifier in January 2023, engineered to differentiate between texts generated by humans and those generated by AI. However, this tool was discontinued in July 2023 due to its low accuracy[1], indicating the inherent difficulties in developing effective GPT detectors. Furthermore, current GenAI detectors demonstrate bias against non-native English speakers for educational contexts like writing argumentative essays (Liang et al., 2023). As a result, these detectors face at least two major challenges, accuracy and fairness, indicating the need for improvements in model performance, especially in educational areas.

In this work, we propose that, by identifying creativity features that distinguish the hu-

[1] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

man from the GenAI *writing process* and that make the two inherently different, irrespective of a human's native language or the sophistication of the GenAI model, one can then exploit such features to build an effective classifier.

Thus, let us start with an analysis of a fundamental difference between human-written and AI-generated content, that is, their purpose. AI-generated content relies on *prediction*: Generative Pre-trained Transformers, or GPTs, are powerful neural networks predicting the most likely (however long) set of words to appear after a textual input or prompt. More generally, GenAI predicts language in terms of high dimensional probability distributions over all the possible words (tokens) the network was trained on. The result is, naturally, a series of words that are somehow "expected", in the sense that they are, statistically speaking, likely to be seen. One of the major issues of GenAI is, in fact, the repetition problem, or the generation of text with redundant rather than new, "unexpected" segments (Fan, Lewis, and Dauphin, 2018; Holtzman et al., 2020)[2]. In contrast, the purpose of human-written content is to put into words an inner thought or feeling, often driven by *exploration*, aiming not just to convey information but also to reflect on complex ideas, express divergent thinking, and introduce new perspectives.

Therefore, to build an effective GenAI detection tool, we propose the use of a set of features, motivated by the creativity literature, which captures content divergence – specifically, semantic dissimilarities among sentences – which stem from the inherent differences in writing purpose between humans and GenAI. We then design a GenAI detection framework

[2] This repetition problem refers to the generation of outputs with redundant segments. Fu et al. (2021) illustrate this with an example where an AI produces a sentence with unnecessary repetition: "Though it is still unfinished, but I like it but I like it but I like ...". The root cause of this problem, however, remains unidentified (Welleck et al., 2019).

by incorporating such creativity-related features into any current GenAI detector. Lastly, we evaluate the proposed GenAI detector focusing on performance improvements in both accuracy and fairness.

Prior literature on divergent thinking and creativity introduced an objective way to assess human verbal creativity based on the underlying assumption that creative individuals choose words with larger semantic distances between them (Olson et al., 2021). In one study, participants were asked to generate seven unrelated nouns, based on which a Divergent Association Task score, or **DAT**, was calculated at the word level, which represents the semantic distance between each pair of words (Olson et al., 2021). More creative people (assessed by standard measures of creativity) generated nouns with higher DAT. Building on this stream of work, we propose extending the DAT metric from a word level to a sentence-level analysis, conceptualizing a new metric to capture nuances in sentence dissimilarity within the text. We build on prior work on creativity and equip the concept with a sentence embedding[3] (Le and Mikolov, 2014) measure. We introduce the DAT score at the sentence level, **DAT(Sent)**, as a metric to proxy the degree of semantic similarity among sentences within a document. Specifically, DAT(Sent) calculates the semantic distances among all possible pairs of sentences within a document through the sentence-embeddings approach. A lower DAT(Sent) indicates smaller distances among sentence embeddings, or greater similarity among sentences in a document, while a higher DAT(Sent) indicates higher divergence. Unlike previous research that compares entire documents to a median or centroid reference

[3] "Word embeddings are a way of representing words as vectors in a multi-dimensional space, where the distance and direction between vectors reflect the similarity and relationships among the corresponding words" (https://www.ibm.com/topics/word-embeddings). Compared to word embeddings, sentence embeddings provide a more comprehensive representation of the semantic meaning of text (Le and Mikolov, 2014).

within a group, which can change with the addition of new samples (Doshi and Hauser, 2024), DAT(Sent) is a more stable metric that is independent of external samples. Following the idea of capturing content divergence through sentence embeddings, we propose three additional features to measure sentence divergence within each document. These include one global[4] content divergence metric at the sentence level, Variance(Sent), and two local[5] content divergence metrics that track dynamic sentence similarities, Diff(Sent) and Diff2(Sent). Beyond semantic divergence among sentences, we construct a metric, DAT(Word), that captures semantic divergence at the word level by generalizing the DAT metric (Olson et al., 2021) (see Section Measurements 2.3.1).

Based on these metrics, we design a flexible two-stage classifier, ***GPT-DATector***, as illustrated in Figure 2.1. The first stage involves any existing GenAI detector, followed by a second stage utilizing logistic regression that integrates our proposed metrics, designed to distinguish between human- and AI-generated texts with higher accuracy. Therefore, a flexible and interpretable model inspired by the literature on human creativity, *GPT-DATector* integrates existing GenAI detectors as the first stage model for initial detection with features that capture content divergence at both the sentence and word level. In this study, we choose two state-of-the-art GPT detectors as the first-stage model: the closed-source black-box detector, GPTZero (Tian and Cui, 2023), and the open-source GPT-who (Venkatraman, Uchendu, and Lee, 2024).

Our primary hypothesis is that humans tend to generate text with larger DAT(Sent)

[4] In this study, the term "global" indicates semantic distances among all sentence embeddings in general, considering the set of all sentences.

[5] In this study, the term "local" indicates semantic distances among subsets of sentence embeddings. For instance, the metric of Diff(Sent) considers the semantic distances between the embeddings of each pair of consecutive sentences.

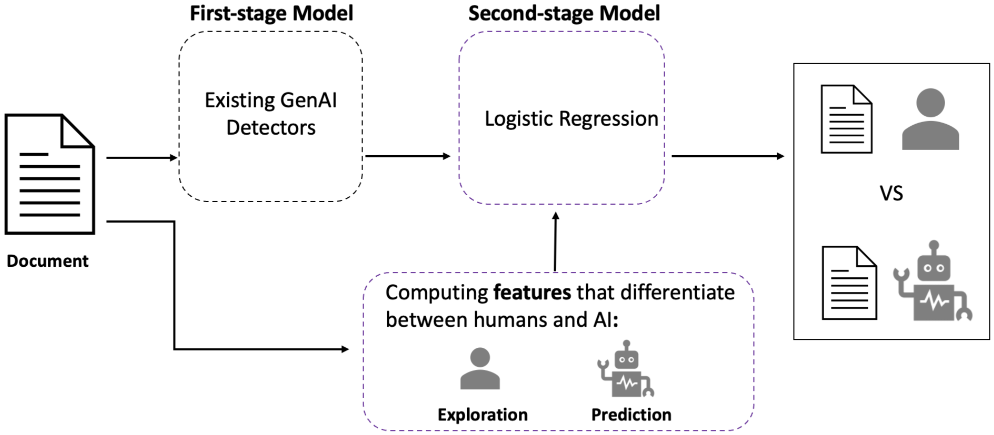


Figure 2.1 Overview of proposed two-stage GPT detector: *GPT-DATector*.

as compared to GenAI; that is, given a group of texts generated by humans and another generated by GenAI, the human-written group would show greater semantic distances among sentences than the AI-generated group. If this is true, incorporating such semantic features at the sentence level can improve the performance of existing GenAI detectors. We test this hypothesis using various writing tasks in traditional school assignments, specifically focusing on argumentative essays and story generation. To construct the samples used for training the GenAI detector, we include essays or stories produced by GenAI alone for each dataset (see Section 2.3.3). After building *GPT-DATector*, we compare both its prediction performance and its bias against non-native English writers to the ones of existing detectors.

We evaluate the prediction performance using two main metrics: accuracy and the area under the receiver operating characteristic curve (AUCROC) (see Section 2.3.4). To further evaluate *GPT-DATector* in mitigating bias against non-native English speakers, we use an additional test set (Liang et al., 2023), which has shown significant bias when other GenAI detectors are applied. We evaluate fairness using two key metrics: Disparate Impact (DI), which captures the ratio of favorable outcomes between underprivileged and privileged groups, and Equal Opportunity Difference (EOD), which measures the difference in true positive rates across these groups. More details can be found in Section 2.3.4.

Our work yields several interesting findings. First, our study shows the presence of significant differences in the distribution of sentence dissimilarity, DAT(Sent), across texts generated by humans and AI. On average, human-written texts exhibit greater DAT(Sent), indicating greater dissimilarity across sentences than AI-generated texts. Interestingly, we find moderate semantic dissimilarity among sentences for texts generated by humans and AI together, given the evidence that the average DAT(Sent) for texts generated by human+AI is larger than in AI-generated stories but smaller than in human-generated stories. In addition, we provide empirical evidence that our proposed *GPT-DATector* outperforms current state-of-the-art detectors, and this result is robust across various writing tasks and types of GenAI. Last, but most importantly, rather than solely achieving larger prediction performance, our results show that *GPT-DATector* can reduce bias in evaluating essays by non-native English speakers when tested on an additional dataset.

This essay is organized as follows. The next section begins with a review of prior literature on the linguistic characteristics of GenAI-generated texts and existing detection methodologies. Section 2.3 introduces the proposed metrics, detection framework, and the dataset used for empirical validation. Section 2.4 presents and interprets the results, demonstrating how and why the proposed metrics effectively distinguish between human- and AI-generated texts, while providing evidence that the framework enhances the prediction accuracy of current GenAI detection models. Finally, Section 2.5 concludes with a summary of findings and discussions of this research.

2.2 Related Literature

In this section, we review two key streams of literature: the linguistic features of texts generated by GenAI and the existing methodologies for detecting such AI-generated content.

2.2.1 Linguistic Characteristics of GenAI-generated Text

Our work is related to the literature on computational linguistics features between GenAI and human-written texts (Beresneva, 2016; Ma et al., 2023). One of those features is perplexity, which measures the average uncertainty of LLM in predicting the next word in a sequence. In general, lower perplexity is observed in AI-generated contexts. For instance, consider the sentence "Hi there, I am an AI", an LLM model is likely to predict the continuation of this sentence with the word "assistant", resulting in a lower perplexity score[6]. In contrast, if the next word is "detector", the perplexity of the sentence would significantly increase, indicating a higher probability of being written by humans. Moreover, Ma et al. (2023) further demonstrates that the lower perplexity observed in AI-generated contexts comes from the training objectives of LLMs. Specifically, these objectives aim to optimize the model's ability to produce text sequences with high probability, leading to a lower perplexity. However, human-written content extends beyond mere generation, but "do things" such as organizing complex information or persuasion.

While perplexity provides a measure of the average uncertainty inherent in LLMs, it does not account for their dynamic characteristics. Specifically, perplexity assesses each word prediction with equal importance, overlooking the bursty nature of language, wherein certain words or phrases are more prevalent in particular contexts. Therefore, the concept of

[6] This example comes from https://support.gptzero.me/hc/en-us/articles/15130070230551

Burstiness is incorporated into GPT detectors to address the patterns of word distribution and occurrence within generated texts. LLM models tend to apply uniform rules for word selection, resulting in lower burstiness. In contrast, a higher value of burstiness is often indicative of human-written content. Thus, burstiness is also considered an important factor among several others in GPT classifiers such as GPTZero.

Recent research includes other linguistic features such as distributions of part-of-speech (POS) tags and named entity (NE) tags for LLM classifiers (Fröhling and Zubiaga, 2021; Crothers, Japkowicz, and Viktor, 2023). Such work is motivated by observed differences in POS tag distributions between human and AI-generated texts (Radford et al., 2019; See et al., 2019). Moreover, See et al. (2019) examines the differences in POS distributions between model-generated and human texts across LLM parameters, discovering that with an increase in K (a parameter related to vocabulary size[7]), the lexical diversity of LLMs reaches that of human-generated text. Specifically, as K nears the vocabulary size, the POS distribution of LLM-generated texts closely fits that of human text, for instance, generating more precise POS categories like Numeral and Proper Noun.

However, those findings are obtained before ChatGPT was developed. Martínez et al. (2024) further evaluates **lexical richness** of texts generated by the latest LLMs such as GPT3.5 and GPT4.0, suggesting that ChatGPT4 produces texts with larger lexical richness than ChatGPT3.5. In this study, they compute the lexical richness based on the total number of words and the count of distinct words following (Tweedie and Baayen, 1998; Van Hout and Vermeer, 2007). Since there exists room in the comparative analysis of lexical

[7] They generate stories using top-k sampling, where the value of K ranges from 1 to vocabulary size. Top-k sampling samples tokens with the highest probabilities until the specified number of tokens is reached.

richness between human-written texts and AI-generated texts (Martínez et al., 2024), our study addresses this gap by introducing a novel methodology for estimating lexical richness. Specifically, we propose a semantic distance metric that employs the word-embedding approach to serve as a proxy for the lexical richness of texts. This innovative metric aims to provide a more nuanced understanding of the textual complexities inherent in human versus AI-generated content.

Our work extends this stream of literature by considering the sentence similarities within the content. Specifically, we propose a metric, DAT(Sent), which is inspired by the divergent thinking creativity literature (Olson et al., 2021). This metric estimates the degree of semantic dissimilarities among sentences within a piece of text. We do this by splitting the text into sentences and mapping each sentence to a high-dimensional space using sentence embeddings, which provides a more nuanced and comprehensive representation of the meaning of text (Le and Mikolov, 2014). Unlike previous research that compares entire documents to a median or centroid reference within a group, which can change by adding new samples (Doshi and Hauser, 2024), DAT(Sent) is a more stable metric independent of external samples.

2.2.2 Current GenAI Detection Methodologies

Our work is also associated with the literature on GenAI detection. Prior research for GenAI detection includes three approaches: 1) classification methods (Solaiman et al., 2019), 2) watermark techniques (Kirchenbauer et al., 2023), and 3) statistical methods (Tian and Cui, 2023; Mitchell et al., 2023). First, classification methods conceptualize the detection of AI-generated text as a binary classification problem. In this approach, a classifier is developed and trained to distinguish between machine-generated and human-generated texts (Crothers et al., 2022; Solaiman et al., 2019). For instance, Desaire et al. (2023) proposed a GPT detector from scratch, employing 20 textual features in conjunction with XGBoost, specifically within the realm of scientific chemistry journals. A typical example is that OpenAI launched a classifier in January 2023, engineered to differentiate between texts generated by humans and those generated by AI, utilizing a RoBERTa-based model (Solaiman et al., 2019). Nevertheless, this tool was discontinued in July 2023 due to its low accuracy[8], suggesting the inherent difficulties in developing highly effective GPT detectors.

The second stream of literature focuses on watermark techniques. Those post-hoc watermarking techniques can be effectively applied to LLMs, which include rule-based approaches (Brassil et al., 1995; Kankanhalli and Hau, 2002) and deep-learning-based strategies (Ueoka, Murawaki, and Kurohashi, 2021; Dai et al., 2022). Kirchenbauer et al. (2023) further introduced a novel approach at inference time, proposing a soft watermarking scheme. This method involves embedding a watermark in each word of a generated sentence through the division of the vocabulary into distinct lists and sampling the next token in a differentiated manner.

The last stream of literature is statistical methods. The general idea is to identify AI-generated text through the analysis of statistical measures such as entropy (Lavergne, Urvoy, and Yvon, 2008) and perplexity (Beresneva, 2016). These methods implement a threshold for the aforementioned statistics to distinguish AI-generated content. Gehrmann, Strobelt, and Rush (2019) introduced the GLTR visualizer, a tool designed to assist humans in the detection of AI-generated text by leveraging entropy, probability, and probability rank for effective detection. Recently, the release of ChatGPT led to the development of two GPT

[8] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

detectors: the closed-source GPTZero (Tian and Cui, 2023) and the open-source DetectGPT (Mitchell et al., 2023). In terms of the closed-sourced GPTZero, it is the leading AI detector with a global user base of over 2.5 million[9], which was trained on a large, diverse corpus of human-written and AI-generated text, with a focus on English prose. GPTZero is capable of detecting AI-generated content at the sentence, paragraph, and document levels through a comprehensive methodology that incorporates linguistic features, including average perplexity[10] and burstiness[11]. In terms of the open-source tool, DetectGPT(Mitchell et al., 2023) uses only log probabilities computed by the model of interest and random perturbations of the passage from another generic pre-trained language model (e.g., T5) (Raffel et al., 2020). Then, AI-text detection is performed by comparing the log probability of the text and its infilled variants.

While there are many detectors, the evolving developments of GenAI present a continuous challenge to effectively differentiate between human- and AI-generated content. For instance, OpenAI launched a classifier in January 2023, engineered to differentiate between texts generated by humans and those generated by AI. However, this tool was discontinued in July 2023 due to its low accuracy[12], indicating the inherent difficulties in developing highly effective GenAI detectors. Therefore, in this study, we design a GenAI detector called *GPT-DATector*, following the idea of using LLMs to combat LLMs (Verspoor, 2024). Specifically, we propose a flexible two-stage framework where the first stage could be any existing GPT

[9] https://gptzero.me/faq

[10] After each word in the text our model develops suggestions of what word is coming next. It checks if our suggestions match what is actually in the text.

[11] The burstiness check analyzes how similar the text is to AI patterns of writing. A human-written document will have changes in style and tone throughout the text, whereas AI content remains similar throughout the document.

[12] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

detector, and the second stage is a logistic regression. In the second stage, we incorporate the output of the stage 1 model and features capturing semantic dissimilarities at both the sentence and word levels. We provide empirical validation to demonstrate that our model achieves higher prediction performance compared to the baseline model.

2.2.3 LLMs and Creativity

Recent research on the creativity of LLMs demonstrates a multi-facet construct with two main aspects, including LLMs' creative capabilities and their capabilities in creative writing tasks. First, existing literature on examining LLMs' creative capabilities largely adopts frameworks developed for assessing human creativity, including subjective and objective assessments. Subjective approaches often rely on human expert judgments to evaluate the novelty and usefulness of LLM-generated outputs across domains such as artistic expression (Crothers, Viktor, and Japkowicz, 2023). Alternative subjective evaluation takes a cognitive perspective, using creativity tasks such as the Alternate Uses Task (AUT), where creativity is measured by the ability to generate unconventional uses for common objects (Guilford, 1964; Summers-Stay, Voss, and Lukin, 2023; Haase and Hanel, 2023). In contrast, objective assessments focus on the underlying structure of semantic memory to evaluate divergent thinking. Among these approaches, the Divergent Association Task (DAT) offers a scalable and valid measure of creativity (Olson et al., 2021) by assessing the semantic distance between words generated by participants (Beaty and Johnson, 2021; Olson et al., 2021). Empirical findings generally demonstrate that LLMs outperform humans in divergent thinking tasks (Bellemare-Pepin et al., 2024; Sun et al., 2025), indicating greater creative capabilities of LLMs.

Furthermore, the literature investigates LLMs in the domain of creative writing, primarily characterizing their function in one of two roles: as autonomous writers of creative content or as collaborative co-authors augmenting human creativity. On one hand, LLMs have been employed to independently generate creative content such as stories, narratives, and design concepts (Bellemare-Pepin et al., 2024; Peeperkorn et al., 2024). On the other hand, LLMs are increasingly used as collaborative writing co-authors, offering suggestions and ideation support for creative tasks (Lee, Liang, and Yang, 2022; Yang et al., 2022; Yuan et al., 2022). This line of research emphasizes the importance of interface design and interaction dynamics in augmenting human creativity (Clark et al., 2018; Bhat et al., 2023).

While LLMs exhibit strong performance in divergent thinking tasks, mixed findings are obtained regarding their effectiveness in complex creative writing contexts. Some studies report that LLMs can surpass human performance in specific creative writing scenarios (Bellemare-Pepin et al., 2024), whereas others demonstrate that LLMs can lag in creative writing (Sun et al., 2025). In this study, we extend the literature on LLMs and creativity by extending the DAT framework from word-level to sentence-level semantic distance. Based on these features, we develop DAT-based GPT detectors and demonstrate that our approach outperforms existing detection methods.

2.3 Methods and Datasets

2.3.1 Measurements

The proposed metric for measuring semantic dissimilarities is inspired by the creativity literature, specifically the DAT score from Olson et al. (2021). They present an objective way to assess human creativity by calculating the semantic distance between pairs of words. The underlying assumption is that creative individuals list words with larger semantic distances between them. Specifically, they ask participants to think of seven unrelated words

$\{word_1, ..., word_7\}$, then map each word into a high-dimensional space via a vector of embeddings. After that, the DAT score is derived as the transformed average of the semantic distances between each pair of words.

Aligned with this literature but equipped with the advantages of sentence embeddings (Le and Mikolov, 2014), in this study, we generalize the DAT score based on sentence embeddings within the document. Specifically, for each document, we segment it into sentences and then map each sentence into a 1536-dimensional vector[13]. Then, for each document with n sentences and their sentence embeddings $\{sent_{eb1}, ..., sent_{ebn}\}$, **DAT(Sent)** is computed as the average cosine distance between all pairs of sentence embeddings, as defined in Equation (2.1). This involves determining the semantic dissimilarities between each pair of sentences using cosine distance and then averaging these distances. To further obtain a measure that ranges from zero to 100, we also multiply the value by 100. A document with a higher DAT(Sent) score indicates greater dissimilarity in semantics among each pair of sentences, and a minimum score of zero indicates that all sentences are identical.

$$DAT(Sent) = \frac{100}{n(n-1)} \sum_{\forall i,j;i \neq j}^{n} {}_{CosineDistance(v_i, v_j)} \tag{2.1}$$

Following the idea of capturing dynamic semantic dissimilarities through sentence embeddings, we further propose three additional semantic features among sentences for each document. (1) $Variance(Sent)$ is calculated as the normalized variance of the sentence embeddings, $Variance(Sent) = \frac{1}{n} \sum_{i=1}^{n} Distance(v_i, \mu)^2$, where μ represents the centroid of all sentence embeddings; (2) $Diff(Sent)$ is calculated as the average of the distances of every two consecutive sentence embeddings v_{i-1} and v_i: $Diff(Sent) = \frac{1}{n-1} \sum_{i=2}^{n} Distance(v_{i-1}, v_i)$;

[13] We adopt OpenAI's state-of-the-art transformer-based embeddings "text-embedding-3-small" model released in January 2024.

and (3) $Diff^2(Sent)$ is calculated as the average of the squared distances of every two consecutive sentence embeddings v_{i-1} and v_i: $Diff^2(Sent) = \frac{1}{n-1}\sum_{i=2}^{n} Distance(v_{i-1}, v_i)^2$. In the following analysis, we refer to these four features related to semantic divergences at the sentence level (including $DAT(Sent)$) as "**Sentence dissimilarities**".

To further capture the semantic divergences at the word level, we propose a $DAT(Word)$ metric based on word embeddings (the orange area in Figure 2.2). The difference in $DAT(Word)$ between our approach and the metric used in the creativity literature (Olson et al., 2021) is that our approach considers all unique words in a long piece of text, rather than a limited number (seven) of nouns generated by humans. Specifically, after eliminating stopwords or punctuation and performing lemmatization, we split each piece of text into a unique set of words (i.e., n unique words), rather than focusing solely on nouns as (Olson et al., 2021) suggested.

2.3.2 Proposed Two-stage GenAI Detector: *GPT-DATector*

The proposed two-stage GenAI detector, *GPT-DATector*, illustrated in Figure 2.1, consists of a first-stage model — flexibly chosen from any existing GenAI detector — and a second-stage logistic regression model. For further clarity, Figure 2.2 presents a more detailed breakdown of the model framework, particularly for the components within the second-stage model. Specifically, in the second-stage model, logistic regression operates using three key feature sets: those derived from the first-stage model, features capturing sentence-level dissimilarities, and features capturing word-level dissimilarities.

First-stage model In this study, we utilize two GenAI detectors as the first-stage models: the open-source GPT-who (Venkatraman, Uchendu, and Lee, 2024) and the closed-source

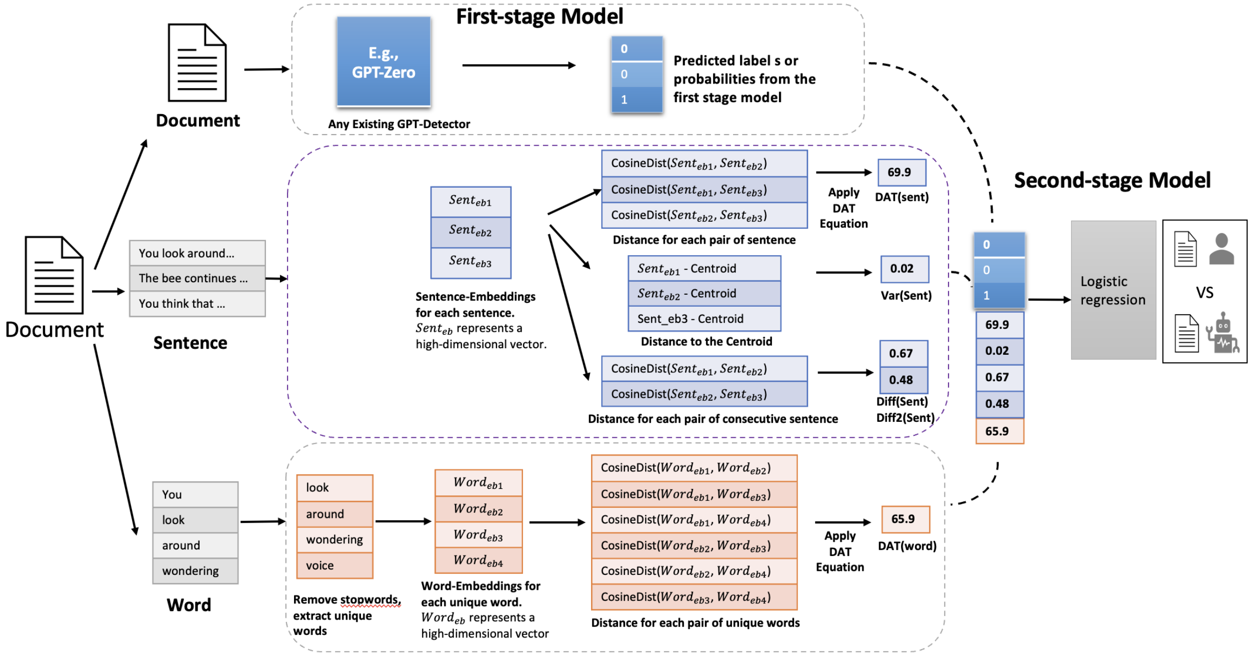

First-stage Model
E.g., GPT-Zero
Any Existing GPT-Detector
0
0
1
Predicted label s or probabilities from the first stage model
Document
Document
You look around...
The bee continues ...
You think that ...
Sentence
Sent_eb1
Sent_eb2
Sent_eb3
Sentence-Embeddings for each sentence. Sent_eb represents a high-dimensional vector.
CosineDist(Sent_eb1, Sent_eb2)
CosineDist(Sent_eb1, Sent_eb3)
CosineDist(Sent_eb2, Sent_eb3)
Distance for each pair of sentence
Apply DAT Equation
69.9
DAT(sent)
Sent_eb1 - Centroid
Sent_eb2 - Centroid
Sent_eb3 - Centroid
Distance to the Centroid
0.02
Var(Sent)
CosineDist(Sent_eb1, Sent_eb2)
CosineDist(Sent_eb2, Sent_eb3)
Distance for each pair of consecutive sentence
0.67
0.48
Diff(Sent)
Diff2(Sent)
Second-stage Model
0
0
1
69.9
0.02
0.67
0.48
65.9
Logistic regression
VS
You
look
around
wondering
Word
look
around
wondering
voice
Remove stopwords, extract unique words
Word_eb1
Word_eb2
Word_eb3
Word_eb4
Word-Embeddings for each unique word. Word_eb represents a high-dimensional vector
CosineDist(Word_eb1, Word_eb2)
CosineDist(Word_eb1, Word_eb3)
CosineDist(Word_eb1, Word_eb4)
CosineDist(Word_eb2, Word_eb3)
CosineDist(Word_eb2, Word_eb4)
CosineDist(Word_eb3, Word_eb4)
Distance for each pair of unique words
Apply DAT Equation
65.9
DAT(word)

Figure 2.2 Proposed two-stage GPT detector: *GPT-DATector.*

(black-box) GPT-Zero (Tian and Cui, 2023).

We use the open-source GPT-who as the first-stage model in our detector, as it has been shown to outperform several state-of-the-art detectors, including GLTR and DetectGPT, by over 20% across more than 10 domains (Venkatraman, Uchendu, and Lee, 2024). GPT-who employs a psycholinguistically motivated Uniform Information Density (UID)-based feature space, grounded in the theoretical assumption that humans generally distribute information evenly in language production, resulting in smaller fluctuations in the distributions of next-word prediction probabilities (Frank and Jaeger, 2008; Mahowald et al., 2013; Xu and Reitter, 2018). Although empirical analysis shows that machine-generated text tends to have more evenly distributed information, GPT-who, based on this UID-based approach, effectively distinguishes human-written from machine-generated text across various tasks, domains, and datasets.

In addition to the open-source GPT-who, we adopt the closed-source (black-box) GPT-

Zero (Tian and Cui, 2023) as the first-stage model due to its high accuracy relative to competing detectors. GPT-Zero employs a multilayered approach encompassing seven modules, including burstiness, perplexity, and an end-to-end deep learning model[14]. This design enables GPT-Zero to capture well-documented characteristics for differentiating human- and AI-generated content. For each text, GPT-Zero provides a predicted label, which classifies the content as human-generated, AI-generated, or mixed (a blend of human and AI elements).

Second-stage model Our objective is to assess whether incorporating creativity-related features, specifically DAT-based metrics, enhances the predictive accuracy of existing GPT detectors such as GPT-who or GPTZero. Consequently, in the second phase of the modeling process, we evaluate model performance by supplementing the outputs (or features) of these detectors with additional DAT-based features. To further examine the marginal effectiveness of distinct features, we configure models with three feature sets: (1) Stage-1 output alone, serving as the baseline without DAT-based features, (2) Stage-1 output combined with DAT-based sentence-level features, and (3) Stage-1 output combined with both DAT-based sentence-level and word-level features.

When conducting the second stage modeling, we implement a logistic regression model to address two-class classification, discerning between human-written and AI-generated documents. In cases involving the CoAuthors dataset, where a "mixed" category (indicating human-AI co-authorship) is present, we apply a multinomial logistic regression model for three-class classification—an extension of logistic regression suited to multi-class scenarios[15].

[14] https://gptzero.me/technology

[15] A multinomial logistic regression modifies the loss function to cross-entropy loss and adjusts the prediction of probability distribution to a multinomial probability distribution, supporting the multi-class classification problems

2.3.3 Datasets

We examine various writing tasks common in traditional educational settings, with a particular focus on argumentative essay composition and prompted story generation. For argumentative essays, two datasets are considered. (1) **ArguTOEFL(GPT)** from ArguGPT (Liu et al., 2023), which contains 1,680 human-written TOEFL essays and 1,635 essays generated by 7 recent GenAIs (including many variants of ChatGPT) using prompts from TOEFL11 (Blanchard et al., 2013); and (2) **HW(GPT3.5)** contains 1,800 human-written argumentative essays from The Hewlett Foundation(HW)[16] and 1,800 AI-generated essays based on the same prompt. Specifically, the 1,800 human-generated essays are written by U.S. Grade 10 students in response to a specific persuasive prompt. The 1,800 AI-generated essays were created by GPT-3.5 with the same prompt. Instead of using "gpt-3.5-turbo" which is chatty, we adopt the version of "gpt-3.5-turbo-instruct", which is much terser and concise[17].

For prompted story generation, two datasets related to the Reddit WritingPrompts[18] are considered: WP (Li et al., 2023), and CoAuthors (Lee, Liang, and Yang, 2022). (3) **WP(GPT)** from (Li et al., 2023) that contains 800 randomly selected human-written stories and 800 AI-generated stories based on a mix of 27 GenAIs. (4) **CoAuthors(GPT3.5)** based on (Lee, Liang, and Yang, 2022), which enables the exploration of differences in the distribution of DAT(Sent) not only between human- and AI-written samples but also within the human+AI generated samples. Specifically, in this dataset, the percentage of AI-generated words is tracked. This allows us to define two types of human-written samples: one

[16] It is noted that we focus on essays in the second set written by U.S. Grade 10 students. https://www.kaggle.com/competitions/asap-aes/data

[17] https://community.openai.com/t/instructgpt-vs-gpt-3-5-turbo/434241

[18] https://www.reddit.com/r/WritingPrompts

where the majority of content is human-written (i.e., less than 15% AI-generated, comprising 226 stories), and another where the content is human+AI blend (i.e., between 15% and 70% AI-generated, comprising 581 stories). Additionally, we include 700 stories generated solely by GPT3.5. (5) Lastly, we construct an additional dataset, **CoAuthors(GPT4)**, which includes the same human-written and human+AI co-created stories, along with 700 stories generated by GPT4[19] alone. Therefore, unlike the first three datasets, which consist of essays or stories written solely by either humans or GenAI, the two CoAuthors datasets — CoAuthors(GPT3.5) and CoAuthors(GPT4) — include an additional category featuring stories co-created by human writers in collaboration with GenAI.

Lastly, to further investigate the performance of *GPT-DATector* in improving fairness, we evaluate its performance using an additional dataset (Liang et al., 2023), which includes essays written by native and non-native English writers. Specifically, after building *GPT-DATector*, we compare both its prediction performance and its bias against non-native English writers to those of existing detectors.

2.3.4 Training Process and Evaluation Metrics

Training process We first split the samples into train and test sets using a 6:4 ratio. During the training process, parameter optimization involves adjusting the predefined hyperparameter: C (inverse of regularization strength) with values $[0.01, 0.1, 1, 10, 100]$. Moreover, we conduct three-fold cross-validation to select the best parameter that maximizes the area under the receiver operating characteristic (AUCROC). Once we find the best parameter, we retrain the model and then use it to make predictions on the test set. Finally, we compute

[19] We use the latest version of “gpt-4-turbo-preview” model, which was trained using data up to Dec 2023. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

prediction performance on the test set. It is noted that we include text documents that contain at least one complete sentence and range in length from 50 to 600 words.

Evaluation metrics We evaluate the predictive performance of *GPT-DATector* using two metrics: (1) accuracy and (2) the area under the receiver operating characteristic curve (AUCROC). For the two-class classification task, detection accuracy measures the model's ability to correctly classify texts as AI-generated or human-written. However, as detection accuracy can vary with threshold selection, AUCROC is used to evaluate performance across all possible thresholds, providing a comprehensive assessment (Mitchell et al., 2023; Krishna et al., 2024). AUCROC represents the probability that the classifier ranks a randomly chosen positive (AI-generated) sample higher than a randomly chosen negative (human-written) sample, thus capturing both precision and recall. This metric offers a robust evaluation method for detector performance across threshold settings (Mitchell et al., 2023).

For the multiclass classification task on the CoAuthors dataset, which includes three categories, we use two generalized versions of AUCROC: macro-average AUC and weighted-average AUC. The macro-average AUC calculates the AUC for each class separately and then averages these values, giving equal weight to each class regardless of its size. In contrast, the weighted-average AUC adjusts for class imbalance by assigning a weight to each class's AUC based on the number of actual instances in that class before averaging.

For fairness evaluation, we employ two commonly used metrics: (1) Disparate Impact (DI), which quantifies the ratio of favorable outcomes between underprivileged and privileged groups. A DI value of 1 indicates equal benefit across groups, values below 1 indicate a bias favoring the privileged group, and values above 1 indicate a bias favoring the underprivileged group. (2) Equal Opportunity Difference (EOD), which measures the difference in true positive rates between these groups. An EOD of 0 denotes equal benefit, values below 0 indicate a bias toward the privileged group, and values above 0 indicate a bias toward the underprivileged group.

2.4 Empirical Results

2.4.1 Human-generated Texts with Larger Sentence-level Semantic Dissimilarities

Our analysis reveals a significant distinction in sentence-level semantic dissimilarities between human- and AI-generated texts (Figure 2.3). Specifically, compared to AI-generated texts, human-written texts exhibit higher semantic dissimilarity between sentences. We find that the DAT(Sent) distribution differs significantly between the two text types. This difference is robust across varying writing tasks and datasets. On average, human-written texts demonstrate a higher DAT(Sent) score (indicated in yellow) than AI-generated texts (indicated in blue).

The fourth and fifth subplots, representing CoAuthors(GPT3.5) and CoAuthors(GPT4), incorporate results from a set of texts co-created by humans and AI. Analysis reveals that the mean DAT(Sent) for these collaboratively generated stories (shown in gray) falls between that of AI-generated stories (in blue) and human-authored stories (in yellow). This intermediate DAT(Sent) score suggests that human-AI collaborative texts exhibit a moderate degree of sentence-level semantic dissimilarity, bridging the characteristics of both human and AI authorship.

To assess whether the observed differences in the empirical cumulative distribution functions between each class pair are statistically significant and directional, we perform a one-

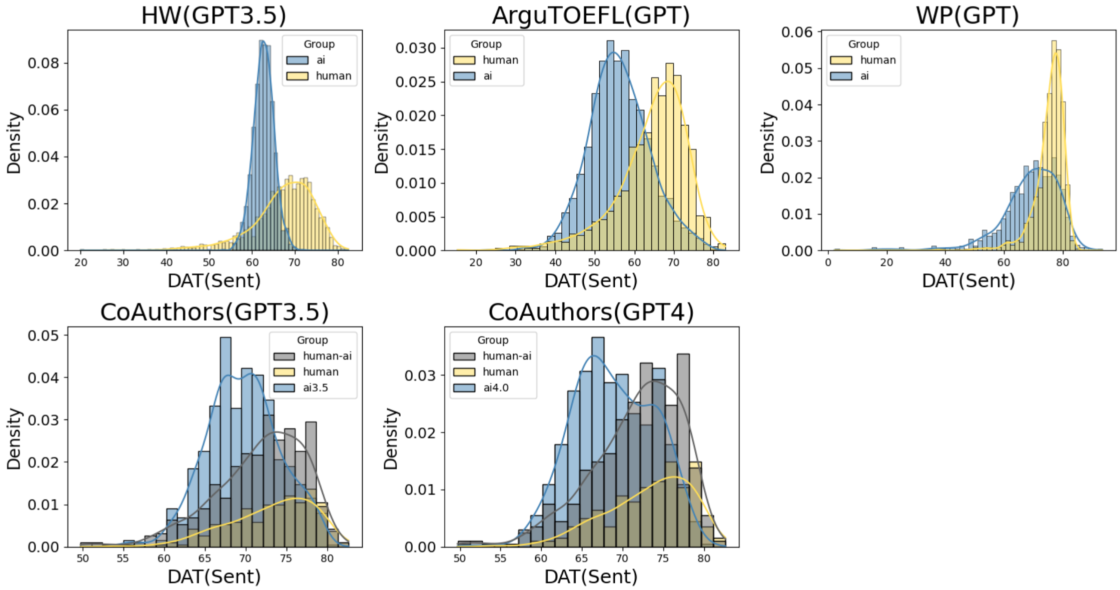


Figure 2.3 Distribution of sentence dissimilarity, DAT(Sent), across human- and AI-generated texts among different datasets.

sided Kolmogorov-Smirnov test. Specifically, the null hypothesis is formulated as H_{0a} : $DAT(Sent)_{human} < DAT(Sent)_{AI}$. For each data source, the null hypothesis H_{0a} is rejected at the 5% significance level, with p-values of $4.12e^{-237}$ (HW(GPT3.5)), $2.93e^{-143}$ (ArguTOEFL(GPT)), $1.54e^{-237}$ (WP(GPT)), $6.88e^{-21}$ (CoAuthors(GPT3.5)) and $3.69e^{-18}$ (CoAuthors(GPT4)), respectively. These findings consistently demonstrate that the DAT(Sent) metric for human-written texts is significantly higher than that for AI-generated texts.

To further investigate trends in human-AI collaborative texts, we conduct two additional one-sided Kolmogorov-Smirnov tests across both CoAuthors(GPT3.5) and CoAuthors(GPT4) datasets. Specifically, we perform two null hypotheses $H_{0b} : DAT(Sent)_{human+AI}$ $< DAT(Sent)_{AI}$; and $H_{0c} : DAT(Sent)_{human} < DAT(Sent)_{human+AI}$. For the CoAuthors(GPT3.5) dataset, we reject H_{0b} at the 5% significance level ($p = 3.92e^{-19}$), indicating human+AI generated texts show significantly higher DAT(Sent) values than AI-generated texts. In contrast, when comparing human+AI texts to human-only texts, we find significantly lower DAT(Sent) values in the collaborative texts, rejecting H_{0c} at the 5% level ($p = 0.005$). Such findings are consistent in the CoAuthors(GPT4) dataset: human+AI co-generated texts have again demonstrated greater DAT(Sent) than AI-generated texts (rejecting H_{0b}, $p = 1.40e^{-19}$) and lower DAT(Sent) than those generated by human only (rejecting H_{0c}, $p = 0.005$).

In summary, our findings reveal clear directional differences in sentence-level semantic dissimilarity, as measured by DAT(Sent), for texts generated solely by AI, solely by humans, and through human-AI collaboration. Specifically, on average, human-generated texts consistently exhibit the highest average DAT(Sent) values, underscoring distinct patterns in semantic structure based on the source of generation. These results suggest that such DAT-related metrics could enhance the detection capabilities of existing GPT classifiers, a potential we further explore in the following section by assessing the impact of integrating these features on classifier performance.

2.4.2 Enhanced Predictive Performance by Leveraging Semantic Dissimilarities among Sentences and Words

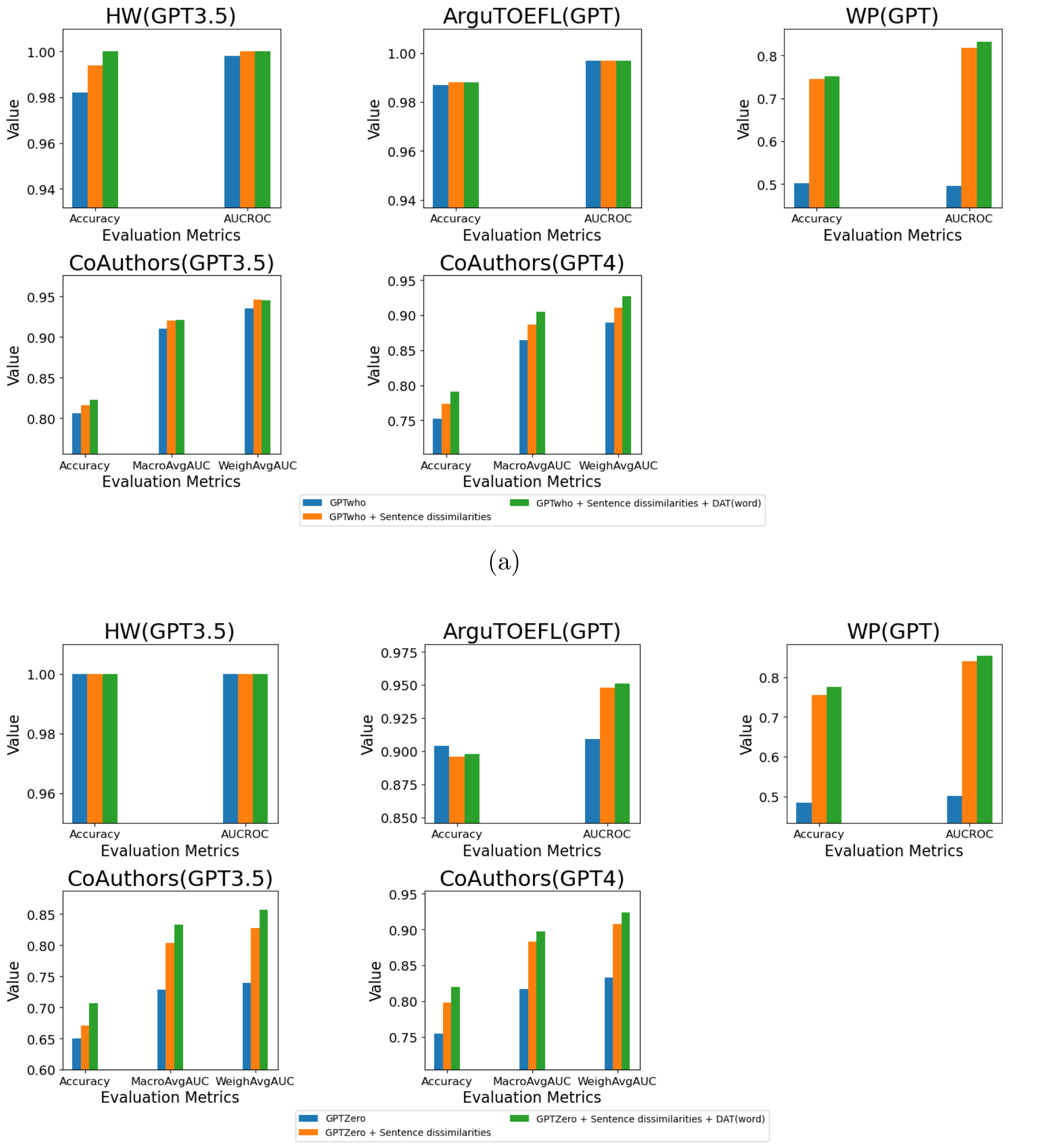
Figure (2.4a) illustrates that our proposed two-stage model, $GPT-DATector$, which integrates features capturing semantic dissimilarities among sentences (in orange), consistently outperforms the baseline stage-1 model (in blue). Specifically, $GPT-DATector$, leveraging the open-source GenAI detector GPT-who (Venkatraman, Uchendu, and Lee, 2024) as its initial stage, was evaluated across multiple datasets with distinct feature sets. Prediction performance on test datasets is reported (see Methods, Section 2.3.2). Results show that incorporating sentence-dissimilarity features—namely DAT(Sent), Var(Sent), Diff(Sent), and Diff2(Sent)—significantly improves both accuracy and AUCROC compared to the baseline.

The third subplot in Figure 2.4a specifically demonstrates an increase in AUCROC from 0.496 to 0.818 on the WP(GPT) dataset with $GPT - DATector$.

In addition, we consistently observe enhanced performance in the three-class classification task (distinguishing between human-generated, AI-generated, or collaboratively human+AI-generated texts) using the CoAuthors dataset. The fourth subplot in Figure 2.4a illustrates that, compared to the baseline model (in blue), $GPT-DATector$ with sentence-level semantic dissimilarity features (in orange) improves the macro-average AUC from 0.910 to 0.920 and the weighted-average AUC from 0.935 to 0.946 on the CoAuthors(GPT3.5) dataset. Similarly, the fifth subplot in Figure 2.4a reveals a similar improvement on the CoAuthors(GPT4.0) dataset, where macro-average AUC rises from 0.864 to 0.887 and weighted-average AUC increases from 0.890 to 0.911. Taken together, these consistent enhancements across varying datasets address the robustness of our approach.

Furthermore, augmenting the model with DAT(Word) on top of sentence-level features, which captures semantic dissimilarities at both word and sentence levels (the green bars in Figure 2.4a, achieves the highest accuracy and AUCROC. This suggests that features encoding content divergence across multiple linguistic levels—sentence and word—substantially enhance the model's capability to distinguish between human- and AI-generated texts. This finding aligns with recent literature advocating for the use of LLMs (i.e., in our study, using sentence- or word-embeddings) to counter LLMs (i.e., in our study, differentiate texts written by GenAI from texts written by humans) (Verspoor, 2024).

Lastly, to demonstrate the generalizability and flexibility of $GPT - DATector$, we replace the stage 1 model, the open-sourced GPT-who, with the closed-sourced GPTZero. Comparable findings are observed in Figure 2.4b. In the second subplot, analyzing the



(a)

(b)

Figure 2.4 Test set performance (Accuracy and AUCROC) across different datasets, where the stage 1 model is the open-sourced GPT-who (**Fig. 4(a)**) or the closed-sourced GPTZero (**Fig. 4(b)**), and the stage 2 model is the (multinomial) Logistic regression. Note: For CoAuthors(GPT3.5) and CoAuthors(GPT4) in a three-class classification scenario (human, AI, and human+AI), we implement multinomial logistic regression, replacing logistic regression as the stage-2 model. Additionally, we substitute the AUCROC evaluation metric with two complementary metrics, MacroAverageAUC and WeightedAverageAUC, for more comprehensive performance assessment.

ArguTOEFL(GPT) dataset[20], $GPT-DATector$ demonstrates enhanced AUCROC values, highlighting the superior predictive capacity of $GPT-DATector$, particularly when integrating semantic divergence across both sentence and word levels (in green), as shown in the second subplot of Figure 2.4b.

Therefore, the proposed $GPT-DATector$, integrating a comprehensive set of features that capture semantic dissimilarities at both sentence and word levels, consistently outperforms baseline models (i.e., existing GenAI detectors such as GPT-who or GPTZero) across varying datasets. This enhanced prediction performance highlights the effectiveness of incorporating multi-level semantic features in advancing the misuse of GenAI, facilitating our understanding of distinguishing characteristics between human- and AI-generated texts.

2.4.3 Reducing Bias against Non-Native English Writers

So far, we have demonstrated improved predictive performance of $GPT-DATector$. In the following section, we investigate whether $GPT-DATector$ that incorporates semantic dissimilarity features at both sentence and word levels, can improve fairness. Our analysis uses human-generated texts[21] from two groups: a privileged group (native English speakers) and an underprivileged group (non-native English speakers).

In Section 2.3.1, we propose DAT(Sent) by hypothesizing that, unlike GenAI, human

[20] We find a slight decrease in accuracy when incorporating different feature sets—sentence-level features alone (in orange) or both sentence- and word-level features (in green)—compared to the baseline model (blue). However, accuracy, limited by a fixed probabilistic threshold of 0.5, inadequately captures the model's discriminative ability due to its sensitivity to class imbalance. When evaluated with AUCROC, a more robust performance metric, the model shows a substantial improvement.

[21] We assess $GPT-DATector$ on a dataset (Liang et al., 2023) containing 91 TOEFL essays by non-native English speakers and 88 essays by U.S. native speakers, which demonstrates that existing GPT detectors exhibit bias against non-native English writers by disproportionately misclassifying their work as AI-generated.

writers tend to display greater exploratory, divergent thinking characteristics in language production (Figure 2.3). To further assess the effectiveness of DAT(Sent) in mitigating bias across human-written texts from both native and non-native English speakers, we expect similar distributions of DAT(Sent) for each group, reflecting the shared human origin of both sets. Ideally, the distribution of $DAT(Sent)_{NativeSpeaker}$ should align closely with $DAT(Sent)_{NonNativeSpeaker}$ since both groups are generated by humans. But practically, our expectation is for DAT(Sent) distributions across these groups to exhibit greater overlaps with one another than those of features typically used in existing GenAI detectors (i.e., the key feature used in GPT-who, the variance in uniform information density UID(Var)).

Figure 2.5 supports our hypothesis. First, the left subplot demonstrates that native and non-native English speakers exhibit comparable DAT(Sent) distributions, shown through kernel density estimation. Specifically, the distributions of DAT(Sent) for native English speakers (in yellow) and non-native speakers (in orange) reveal substantial overlap, indicating similar trends in sentence-level semantic dissimilarity across both groups. This overlap aligns with our expectation, as both groups consist of human writers and thus reflect common divergent thinking capabilities in language generation, irrespective of native language status (i.e., native English speaker or not).

In contrast, the right subplot in Figure 2.5 represents the distribution of UID(Var)[22]—a key feature used in GPT-who[23], our baseline model—indicating its inherent bias within GPT-

[22] Specifically, human-written text shows greater variance in *surprisal* for next-word predictions compared to AI-generated text (Venkatraman, Uchendu, and Lee, 2024). It is noted that the term "*surprisal*" of words refers to how unexpected a word is within a given context from (Hale, 2001; Xu and Reitter, 2018) compared to AI-generated text; less predictable words have larger surprisal, while highly predictable words have less information.

[23] In the analysis regarding fairness performance, we focus on open-source detectors like GPT-who. The reason is that, compared to black-box detectors such as GPTZero which lack

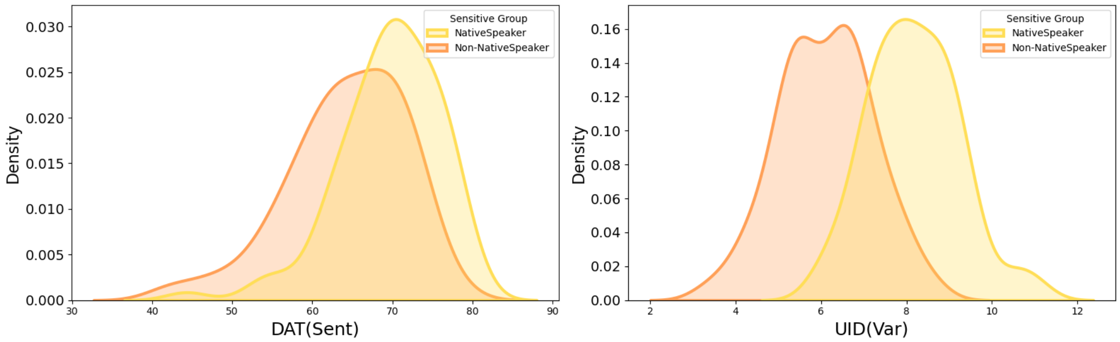


Figure 2.5 Model-free evidence of DAT(Sent) in mitigating bias against non-native English speakers on an additional dataset (Liang et al., 2023). Specifically, two sub-figures represent the kernel density estimation of DAT(Sent) and UID(Var) distributions, respectively. UID(Var), a key feature used in the baseline model GPT-who, quantifies *surprisal* in next-word prediction (Hale, 2001; Xu and Reitter, 2018). In contrast, our proposed feature, DAT(Sent), reflects semantic dissimilarities between sentences.

who. While GPT-who leverages UID-related features to differentiate between human and AI-generated content, such UID-related features reveal distinct distributions between native and non-native English speakers, despite both groups being human-written. Such findings align well with previous research that non-native English writers generally demonstrate lower linguistic variability, including reduced lexical richness (Laufer and Nation, 1995), lexical diversity (Jarvis, 2002), and syntactic complexity (Lu, 2011).

We further compute a Kolmogorov-Smirnov (KS) test statistic[24] to investigate whether the distribution of DAT(Sent) exhibits statistically greater similarity between native and non-native English speakers compared to UID(Var). A KS statistic of 0.33 for DAT(Sent) and 0.70 for UID(Var). The lower KS statistic for DAT(Sent) signifies a significantly higher

explainability and interoperability, open-source detectors like GPT-who include key features that can be analyzed and interpreted.

[24] In this case, we conduct the two-sided Kolmogorov-Smirnov test. Specifically, H_{0d} : $DAT(Sent)_{NativeSpeaker} = DAT(Sent)_{NonNativeSpeaker}$; and H_{0e} : $UID(Var)_{NativeSpeaker} = UID(Var)_{NonNativeSpeaker}$. The KS statistic quantifies the maximum difference between the cumulative distribution functions of two groups, with smaller KS statistics indicating a smaller distance or closer similarity between distributions.

distributional overlap between essays written by native and non-native English speakers in DAT(Sent) than in UID(Var), supporting our hypothesis of greater similarity in DAT(Sent) for human-generated texts.

Table 2.1 Evaluating *GPT-DATector* on additional test set (Fairness-related metrics).

Fairness Metric	Training Samples: HW(GPT3.5)		Training Samples: HW(GPT3.5) + ArguTOEFL(GPT)	
	GPT-who (baseline)	Our approach (*GPT-DATector* with all features[2])	GPT-who (baseline)	Our approach (*GPT-DATector* with all features)
Disparate Impact (DI)	0.58	1.00	0.69	0.85
Equal Opportunity Difference (EOD)	-0.43	-0.02	-0.32	-0.17
%HumanMisclassified AsAI	underprivileged -0.43 privileged = 0.0	underprivileged -0.02 privileged = 0.0	underprivileged -0.33 privileged = 0.0	underprivileged -0.17 privileged = 0.0

Note: (1) "Underprivileged" refers to essays written by non-native English speakers, and "Privileged" refers to essays written by native English speakers. (2) "All features" represent the full set of proposed features, including four semantic dissimilarities among sentences (DAT(Sent), Var(Sent), Diff(Sent), and Diff2(Sent)), and one semantic dissimilarity among words (DAT(Word)).

Lastly, to assess whether *GPT-DATector* could improve fairness, we train the model on two argumentative datasets, HW(GPT3.5) and ArguTOEFL(GPT)[25], and subsequently apply to an additional dataset Liang et al. (2023) to classify essays as AI- or human-generated. Table 2.1 reports the results, suggesting the model's effectiveness in enhancing fairness. When training on one argumentative dataset, HW(GPT3.5) only, we find that *GPT-DATector* with the full set of proposed features improves DI from 0.58 to 1 (where 1 indicates equal benefit) and reduces EOD from -0.43 to -0.02 (where 0 indicates equal benefit), compared to the baseline model (GPT-who). Similar findings are observed when the model is trained on two argumentative datasets together, HW(GPT3.5) and ArguTOEFL(GPT).

[25] We do this because the additional dataset Liang et al. (2023) is related to the argumentative writing task, rather than prompted story generations.

Therefore, our proposed *GPT-DATector*, incorporating sentence- and word-level semantic similarity, effectively promotes fairness between native and non-native English writers by reducing bias against the latter. These results demonstrate that *GPT-DATector* has significant potential for application in reducing bias within educational settings such as college admissions.

2.5 Discussion and Conclusion

2.5.1 Findings

GenAI is undergoing rapid evolution, with the continuous emergence of newly developed detection tools. Aligned with the idea of using LLMs to combat LLMs (Verspoor, 2024) and divergent thinking creativity literature (Olson et al., 2021), this study proposes DAT(Sent) as a metric to proxy semantic dissimilarities within the text. We show that, on average, human-generated contents have a larger DAT(Sent) than AI-generated texts across different writing tasks and datasets. Moreover, we design a GenAI detector, *GPT-DATector*, that incorporates a set of features that capture sentence-level and word-level semantic dissimilarity. Empirical validations demonstrate that our proposed *GPT-DATector* outperforms state-of-the-art models like GPTZero and GPT-who in terms of predictive performance. Most importantly, we find that *GPT-DATector* can reduce bias against non-native English speakers, as evidenced by its application to an additional dataset.

2.5.2 Contribution

Our work makes several contributions as below. First, our study contributes to the computational linguistics literature by advancing methods to differentiate between GenAI-generated and human-generated texts, integrating insights from divergent thinking studies. While prior research has predominantly studied metrics such as perplexity (i.e., a measure of uncertainty in predicting word occurrences within a model) (Wallach et al., 2009; Beresneva, 2016), our study aligns with this stream of literature by introducing a novel set of DAT-related metrics leveraging state-of-the-art embedding techniques. We demonstrate that the distribution of DAT(Sent) effectively distinguishes AI-generated texts from human-written ones, providing deeper insights into the linguistic characteristics of GenAI outputs and complementing emerging research on using LLMs to address challenges posed by LLMs (Verspoor, 2024). Additionally, unlike previous approaches that evaluate sentence- or document-level similarity by comparing entire documents to a median or central reference (Doshi and Hauser, 2024; Zhou and Lee, 2024), our proposed metrics, can capture the degree of semantic repetition within a document, offering a more robust and insightful evaluation of LLM-generated texts.

Moreover, our study makes a significant contribution to the GenAI detection literature by introducing *GPT-DATector*, a framework grounded in the fundamental distinctions between human and AI-generated content. While revolutionary in GenAI's capabilities, it poses significant risks such as deepfakes, synthetic identities, and highly precise misinformation, highlighting their dual-edged nature (Ferrara, 2024). We demonstrate that the proposed two-stage detection model, *GPT-DATector*, not only achieves higher accuracy but also mitigates bias compared to existing state-of-the-art models.

Lastly, the proposed two-stage framework has broad applicability as it can be integrated with a range of state-of-the-art GPT detection methods, where the first-stage model can be substituted with any existing GenAI detector (either the black-box or open-source model). Therefore, our proposed *GPT-DATector* offers substantial practical implications by providing a reliable tool for identifying the origin of texts, whether human-written, AI-generated, or a combination of both.

2.5.3 Limitations and Future Work

While our dataset encompasses a range of educational writing tasks such as argumentative and creative writing, it remains limited in scope. Future research could enhance generalizability by incorporating additional writing tasks that are prevalent in educational contexts, such as reflective essays and explanatory writing. Furthermore, improving the effectiveness and reliability of GPT detectors necessitates a comprehensive understanding of human writing styles. This, in turn, requires a sufficiently large and diverse set of human-written samples to ensure the validity of comparisons. Thus, future work should also consider strategies to enhance the representativeness and the quality of human-written texts, as this is essential for accurately evaluating and strengthening the generalizability of GPT detection models.

Another limitation of this work is the absence of extensive prompt engineering, which has been shown to significantly influence the quality of GenAI outputs (Nori et al., 2023; Zamfirescu-Pereira et al., 2023). Our design employed a single prompt to generate a typical real-world scenario in which non-expert users interact with LLMs and expect coherent, high-quality responses without iterative refinement (Zamfirescu-Pereira et al., 2023; Sun et al., 2025). Repeated prompting or interactive refinement may produce more human-like text, thereby increasing the difficulty for GenAI detection models. Future research should explore how varying levels of prompt engineering affect both the quality of AI-generated content and the robustness of GPT detection models.

As GenAIs continue to evolve, future LLMs are expected to exhibit significantly enhanced capabilities in interpreting user intent and producing more sophisticated outputs (Acar et al., 2024). Consequently, the effectiveness of GPT detectors may vary depending on the specific model generating the text. The increased linguistic and contextual fluency of newer models poses challenges for detection approaches calibrated to earlier-generation outputs. Therefore, future research should evaluate the performance of the proposed *GPT-DATector* using texts generated by more advanced LLMs.

2.5.4 Conclusion

In this study, we address the challenge of distinguishing AI-generated texts from those written by humans. Differ from prior work, we begin with a critical observation often overlooked: the repetition problem in GenAIs, wherein AI-generated responses contain semantically redundant segments across sentences (Fan, Lewis, and Dauphin, 2018; Holtzman et al., 2020; Fu et al., 2021). Motivated by this phenomenon, and informed by the idea of leveraging LLMs to detect LLM outputs (Verspoor, 2024) and insights from divergent thinking and creativity research (Olson et al., 2021), we develop a novel sentence-level semantic metric that quantifies the degree of semantic repetition across sentences. Empirically, we find robust evidence that human-written texts exhibit significantly greater sentence-level semantic dissimilarity than AI-generated ones. These findings reveal a fundamental structural divergence between human and machine-written content. Importantly, our proposed GPT detector, *GPT-DATector*, not only improves the detection of AI-generated text but also mitigates bias against non-native English writers, which is a key concern in the fair evaluation of language quality and authorship.

The rapid advancement of GenAI has narrowed the gap between human and AI capabilities, making it increasingly difficult to distinguish between human- and AI-generated texts. However, beyond detection challenges, GenAI may also be reshaping human distinctiveness through continued interaction in the long run. Recent evidence suggests that while LLMs can temporarily enhance creativity, they may hinder users' independent creative thinking when users discontinue AI assistance (Kumar et al., 2025). Furthermore, reliance on GenAI can impair core human capabilities in SAT essay writing. Users assisted by ChatGPT show the lowest brain engagement and underperform across neural, linguistic, and behavioral dimensions (Kosmyna et al., 2025). As human writing converges with AI output, the foundational assumptions of AI-detection models that rely on divergence between human and AI may no longer hold. Therefore, beyond developing the most effective GenAI detection models, it is essential to examine the evolving role of humans in the future of work, particularly in educational contexts. So that as educators we can design information systems that preserve and strengthen human cognitive and creative capabilities rather than diminishing them through overreliance on AI.

BIBLIOGRAPHY

Acar, O.A., A. Tuncdogan, D. van Knippenberg, and K.R. Lakhani. 2024. "Collective creativity and innovation: An interdisciplinary review, integration, and research agenda." *Journal of Management* 50:2119–2151.

Amabile, T.M. 2018. *Creativity in context: Update to the social psychology of creativity*. Routledge.

Anthony, C., B.A. Bechky, and A.L. Fayard. 2023. ""Collaborating" with AI: Taking a system view to explore the future of work." *Organization Science* 34:1672–1694.

Arnold, K.C., K. Chauncey, and K.Z. Gajos. 2020. "Predictive text encourages predictable writing." In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. pp. 128–138.

Arriagada, L. 2020. "CG-Art: Demystifying the anthropocentric bias of artistic creativity." *Connection Science* 32:398–405.

Bauer, K., M. von Zahn, and O. Hinz. 2023. "Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing." *Information systems research* 34:1582–1602.

Beaty, R.E., and D.R. Johnson. 2021. "Automating creativity assessment with SemDis: An open platform for computing semantic distance." *Behavior research methods* 53:757–780.

Bellemare-Pepin, A., F. Lespinasse, P. Thölke, Y. Harel, K. Mathewson, J.A. Olson, Y. Bengio, and K. Jerbi. 2024. "Divergent Creativity in Humans and Large Language Models." *arXiv preprint arXiv:2405.13012*, pp. .

Benedek, M., M. Karstendiek, S.M. Ceh, R.H. Grabner, G. Krammer, I. Lebuda, P.J. Silvia, K.N. Cotter, Y. Li, W. Hu, et al. 2021. "Creativity myths: Prevalence and correlates of misconceptions on creativity." *Personality and Individual Differences* 182:111068.

Beresneva, D. 2016. "Computer-generated text detection using machine learning: A systematic review." In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*. Springer, pp. 421–426.

Bhat, A., S. Agashe, P. Oberoi, N. Mohile, R. Jangir, and A. Joshi. 2023. "Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing." In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. pp. 436–452.

Blanchard, D., J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. 2013. "TOEFL11: A corpus of non-native English." *ETS Research Report Series* 2013:i–15.

Boussioux, L., J.N. Lane, M. Zhang, V. Jacimovic, and K.R. Lakhani. 2024. "The crowdless future? Generative AI and creative problem-solving." *Organization Science* 35:1589–1607.

Brassil, J.T., S. Low, N.F. Maxemchuk, and L. O'Gorman. 1995. "Electronic marking and identification techniques to discourage document copying." *IEEE Journal on Selected Areas in Communications* 13:1495–1504.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. "Language models are few-shot learners." *Advances in neural information processing systems* 33:1877–1901.

Brynjolfsson, E. 2023. "The turing trap: The promise & peril of human-like artificial intelligence." In *Augmented education in the global age*. Routledge, pp. 103–116.

Brynjolfsson, E., D. Li, and L. Raymond. 2025. "Generative AI at work." *The Quarterly Journal of Economics*, pp. qjae044.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, et al. 2023. "Sparks of artificial general intelligence: Early experiments with GPT-4." *arXiv preprint arXiv:2303.12712*, pp. .

Bunjak, A., M. Černe, and A. Popovič. 2021. "Absorbed in technology but digitally overloaded: Interplay effects on gig workers' burnout and creativity." *Information & Management* 58:103533.

Burton, J.W., M.K. Stein, and T.B. Jensen. 2020. "A systematic review of algorithm aversion in augmented decision making." *Journal of behavioral decision making* 33:220–239.

Buschek, D., M. Zürn, and M. Eiband. 2021. "The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. pp. 1–13.

Castelo, N., M.W. Bos, and D.R. Lehmann. 2019. "Task-dependent algorithm aversion." *Journal of Marketing Research* 56:809–825.

Chakrabarty, T., P. Laban, D. Agarwal, S. Muresan, and C.S. Wu. 2024. "Art or artifice? large language models and the false promise of creativity." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. pp. 1–34.

Chen, Z., and J. Chan. 2024. "Large language model in creative work: The role of collabo ration modality and user expertise." *Management Science* 70:9101–9117.

Clark, E., A.S. Ross, C. Tan, Y. Ji, and N.A. Smith. 2018. "Creative writing with a machine in the loop: Case studies on slogans and stories." In *23rd International Conference on Intelligent User Interfaces*. pp. 329–340.

Clerwall, C. 2017. "Enter the robot journalist: Users' perceptions of automated content." In *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty*. Routledge, pp. 165–177.

Commerford, B.P., S.A. Dennis, J.R. Joe, and J.W. Ulla. 2022. "Man versus machine: Complex estimates and auditor reliance on artificial intelligence." *Journal of Accounting Research* 60:171–201.

Crothers, E., N. Japkowicz, H. Viktor, and P. Branco. 2022. "Adversarial robustness of neural-statistical features in detection of generative transformers." In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Crothers, E., N. Japkowicz, and H.L. Viktor. 2023. "Machine-generated text: A comprehensive survey of threat models and detection methods." *IEEE Access*, pp. .

Crothers, E., H. Viktor, and N. Japkowicz. 2023. “In BLOOM: Creativity and Affinity in Artificial Lyrics and Art.” *arXiv preprint arXiv:2301.05402*, pp. .

Dai, L., J. Mao, X. Fan, and X. Zhou. 2022. “Deephider: A multi-module and invisibility watermarking scheme for language model.” *arXiv preprint arXiv:2208.04676*, pp. .

Damon, W., and E. Phelps. 1989. “Critical distinctions among three approaches to peer education.” *International journal of educational research* 13:9–19.

Dathathri, S., A. See, S. Ghaisas, P.S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. 2024. “Scalable watermarking for identifying large language model outputs.” *Nature* 634:818–823.

Desaire, H., A.E. Chua, M.G. Kim, and D. Hua. 2023. “Accurately detecting AI text when ChatGPT is told to write like a chemist.” *Cell Reports Physical Science* 4.

Dietvorst, B.J., J.P. Simmons, and C. Massey. 2015. “Algorithm aversion: people erroneously avoid algorithms after seeing them err.” *Journal of Experimental Psychology: General* 144:114.

—. 2018. “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them.” *Management science* 64:1155–1170.

Doshi, A.R., and O.P. Hauser. 2024. “Generative AI enhances individual creativity but reduces the collective diversity of novel content.” *Science Advances* 10:eadn5290.

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.

Eining, M.M., D.R. Jones, J.K. Loebbecke, et al. 1997. “Reliance on decision aids: An examination of auditors’ assessment of management fraud.” *Auditing: A Journal of practice & theory* 16.

Emig, J. 1971. “The composing processes of twelfth graders.” *National Council of teachers of English research report*, pp. .

Fan, A., M. Lewis, and Y. Dauphin. 2018. “Hierarchical neural story generation.” *arXiv preprint arXiv:1805.04833*, pp. .

Ferrara, E. 2024. “GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models.” *Journal of Computational Social Science* 7:549–569.

Frank, A.F., and T.F. Jaeger. 2008. “Speaking rationally: Uniform information density as an optimal strategy for language production.” In *Proceedings of the annual meeting of the cognitive science society*. vol. 30.

Fröhling, L., and A. Zubiaga. 2021. “Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover.” *PeerJ Computer Science* 7:e443.

Fu, Z., W. Lam, A.M.C. So, and B. Shi. 2021. “A theoretical analysis of the repetition problem in text generation.” In *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 12848–12856.

Fügener, A., J. Grahl, A. Gupta, and W. Ketter. 2022. "Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation." *Information Systems Research* 33:678–696.

Ge, R., Z. Zheng, X. Tian, and L. Liao. 2021. "Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending." *Information Systems Research* 32:774–785.

Gehrmann, S., H. Strobelt, and A.M. Rush. 2019. "Gltr: Statistical detection and visualization of generated text." *arXiv preprint arXiv:1906.04043*, pp. .

Ghaffari, S., B. Yousefimehr, and M. Ghatee. 2024. "Generative-AI in E-Commerce: Use-Cases and Implementations." In *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*. IEEE, pp. 1–5.

Gnewuch, U., S. Morana, O. Hinz, R. Kellner, and A. Maedche. 2024. "More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents." *Information Systems Research* 35:936–955.

Goffman, E. 2017. *Interaction ritual: Essays in face-to-face behavior*. Routledge.

Goncalo, J.A., F.J. Flynn, and S.H. Kim. 2010. "Are two narcissists better than one? The link between narcissism, perceived creativity, and creative performance." *Personality and Social Psychology Bulletin* 36:1484–1495.

Guilford, J.P. 1964. "Some new looks at the nature of creative processes." *Contributions to mathematical psychology. New York: Holt, Rinehart & Winston*, pp. .

Haase, J., and P.H. Hanel. 2023. "Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity." *Journal of Creativity* 33:100066.

Hafenbrädl, S., D. Waeger, J.N. Marewski, and G. Gigerenzer. 2016. "Applied decision making with fast-and-frugal heuristics." *Journal of Applied Research in Memory and Cognition* 5:215–231.

Hale, J. 2001. "A probabilistic Earley parser as a psycholinguistic model." In *Second meeting of the north american chapter of the association for computational linguistics*.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi. 2020. "The curious case of neural text degeneration." *International Conference on Learning Representations (ICLR)*, pp. .

Jakesch, M., A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman. 2023. "Co-writing with opinionated language models affects users' views." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. pp. 1–15.

Jakesch, M., J.T. Hancock, and M. Naaman. 2023. "Human heuristics for AI-generated language are flawed." *Proceedings of the National Academy of Sciences* 120:e2208839120.

Jarrahi, M.H. 2018. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making." *Business horizons* 61:577–586.

Jarvis, S. 2002. "Short texts, best-fitting curves and new measures of lexical diversity." *Language Testing* 19:57–84.

Kane, G.C., A.G. Young, A. Majchrzak, and S. Ransbotham. 2021. “Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants.” *MIS Quarterly* 45:371–396.

Kankanhalli, M.S., and K. Hau. 2002. “Watermarking of electronic text documents.” *Electronic Commerce Research* 2:169–187.

Katz, D.M., M.J. Bommarito, S. Gao, and P. Arredondo. 2024. “GPT-4 passes the bar exam.” *Philosophical Transactions of the Royal Society A* 382:20230254.

Kirchenbauer, J., J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. 2023. “A watermark for large language models.” In *International Conference on Machine Learning*. PMLR, pp. 17061–17084.

Kleinmuntz, B. 1990. “Why we still use our heads instead of formulas: toward an integrative approach.” *Psychological bulletin* 107:296.

Köbis, N., and L.D. Mossink. 2021. “Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry.” *Computers in human behavior* 114:106553.

Koltovskaia, S. 2020. “Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study.” *Assessing Writing* 44:100450.

Kosmyna, N., E. Hauptmann, Y.T. Yuan, J. Situ, X.H. Liao, A.V. Beresnitzky, I. Braunstein, and P. Maes. 2025. “Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task.” *arXiv preprint arXiv:2506.08872*, pp. .

Krishna, K., Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. 2024. “Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.” *Advances in Neural Information Processing Systems* 36.

Kumar, H., J. Vincentius, E. Jordan, and A. Anderson. 2025. “Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking.” In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. pp. 1–18.

Landis, J.R., and G.G. Koch. 1977. “The measurement of observer agreement for categorical data.” *Biometrics*, pp. 159–174.

Laufer, B., and P. Nation. 1995. “Vocabulary size and use: Lexical richness in L2 written production.” *Applied linguistics* 16:307–322.

Lavergne, T., T. Urvoy, and F. Yvon. 2008. “Detecting Fake Content with Relative Entropy Scoring.” *Pan* 8:4.

Le, Q., and T. Mikolov. 2014. “Distributed representations of sentences and documents.” In *International conference on machine learning*. PMLR, pp. 1188–1196.

Lebovitz, S., H. Lifshitz-Assaf, and N. Levina. 2022. “To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis.” *Organization science* 33:126–148.

Lee, M., P. Liang, and Q. Yang. 2022. “Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities.” In *Proceedings of the 2022 CHI conference on human factors in computing systems*. pp. 1–19.

Lee, M., M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X.L. Li, F. Ladhak, F. Rong, et al. 2023. “Evaluating human-language model interaction.” *Transactions on Machine Learning Research*, pp. .

Li, N., H. Zhou, W. Deng, J. Liu, F. Liu, and K. Mikel-Hong. 2024. “When Advanced AI Isn’t Enough: Human Factors as Drivers of Success in Generative AI-Human Collaborations.” *Available at SSRN 4738829*, pp. .

Li, Y., Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang. 2023. “MAGE: Machine-generated Text Detection in the Wild.” *arXiv e-prints*, pp. arXiv–2305.

Liang, W., M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. 2023. “GPT detectors are biased against non-native English writers.” *Patterns* 4.

Liu, Y., Z. Zhang, W. Zhang, S. Yue, X. Zhao, X. Cheng, Y. Zhang, and H. Hu. 2023. “ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models.” *arXiv preprint arXiv:2304.07666*, pp. .

Lockhart, E.N. 2024. “Creativity in the age of AI: the human condition and the limits of machine generation.” *Journal of Cultural Cognitive Science*, pp. 1–6.

Logg, J.M., J.A. Minson, and D.A. Moore. 2019. “Algorithm appreciation: People prefer algorithmic to human judgment.” *Organizational Behavior and Human Decision Processes* 151:90–103.

Lu, X. 2011. “A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers’ language development.” *TESOL quarterly* 45:36–62.

Lu, Y., X. Luo, L. Huang, and D. Wang. 2025. “Can Providing Algorithmic Performance Information Facilitate Humans’ Inventory Ordering Behaviors?” *Information Systems Research*, pp. .

Ma, K., D. Grandi, C. McComb, and K. Goucher-Lambert. 2024. “Exploring the Capabilities of Large Language Models for Generating Diverse Design Solutions.” *arXiv preprint arXiv:2405.02345*, pp. .

Ma, Y., J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu. 2023. “AI vs. human–differentiation analysis of scientific content generation.” *arXiv* 2301.

Mahowald, K., E. Fedorenko, S.T. Piantadosi, and E. Gibson. 2013. “Info/information theory: Speakers choose shorter words in predictive contexts.” *Cognition* 126:313–318.

Majchrzak, A., and M.L. Markus. 2012. “Technology affordances and constraints in management information systems (MIS).” *Encyclopedia of Management Theory,(Ed: E. Kessler), Sage Publications, Forthcoming*, pp. .

Markus, M.L., and M.S. Silver. 2008. “A foundation for the study of IT effects: A new look at DeSanctis and Poole’s concepts of structural features and spirit.” *Journal of the Association for Information systems* 9:5.

Martínez, G., J.A. Hernández, J. Conde, P. Reviriego, and E. Merino. 2024. “Beware of Words: Evaluating the Lexical Richness of Conversational Large Language Models.” *arXiv preprint arXiv:2402.15518*, pp. .

May, R. 1994. *The courage to create*. WW Norton &Company, Inc.

McKee, R. 1997. “Substance, structure, style, and the principles of screenwriting.” *Alba Editorial*, pp. .

Millet, K., F. Buehler, G. Du, and M.D. Kokkoris. 2023. “Defending humankind: Anthropocentric bias in the appreciation of AI art.” *Computers in Human Behavior* 143:107707.

Minaee, S., T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao. 2024. “Large language models: A survey.” *arXiv preprint arXiv:2402.06196*, pp. .

Mirowski, P., K.W. Mathewson, J. Pittman, and R. Evans. 2023. “Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. pp. 1–34.

Mitchell, E., Y. Lee, A. Khazatsky, C.D. Manning, and C. Finn. 2023. “DetectGPT: Zero-shot machine-generated text detection using probability curvature.” In *International Conference on Machine Learning*. PMLR, pp. 24950–24962.

Montazemi, A.R. 1991. “The impact of experience on the design of user interface.” *International journal of man-machine studies* 34:731–749.

Morewedge, C.K. 2022. “Preference for human, not algorithm aversion.” *Trends in cognitive sciences* 26:824–826.

Nori, H., Y.T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, et al. 2023. “Can generalist foundation models outcompete special-purpose tuning? case study in medicine.” *arXiv preprint arXiv:2311.16452*, pp. .

Noy, S., and W. Zhang. 2023. “Experimental evidence on the productivity effects of generative artificial intelligence.” *Science* 381:187–192.

Olson, J.A., J. Nahas, D. Chmoulevitch, S.J. Cropper, and M.E. Webb. 2021. “Naming unrelated words predicts creativity.” *Proceedings of the National Academy of Sciences* 118:e2022340118.

Orbell, J., and R.M. Dawes. 1991. “A “cognitive miser” theory of cooperators advantage.” *American Political Science Review* 85:515–528.

Ornes, S. 2019. “Computers take art in new directions, challenging the meaning of “creativity”.” *Proceedings of the National Academy of Sciences* 116:4760–4763.

Peeperkorn, M., T. Kouwenhoven, D. Brown, and A. Jordanous. 2024. “Is temperature the creativity parameter of large language models?” *arXiv preprint arXiv:2405.00492*, pp. .

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. “Language models are unsupervised multitask learners.” *OpenAI blog* 1:9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu. 2020. “Exploring the limits of transfer learning with a unified text-to-text transformer.” *The Journal of Machine Learning Research* 21:5485–5551.

Ragot, M., N. Martin, and S. Cojean. 2020. “Ai-generated vs. human artworks. a perception bias towards artificial intelligence?” In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. pp. 1–10.

Revilla, E., M.J. Saenz, M. Seifert, and Y. Ma. 2023. "Human–artificial intelligence collaboration in prediction: A field experiment in the retail industry." *Journal of Management Information Systems* 40:1071–1098.

Roemmele, M., and A.S. Gordon. 2018. "Automated assistance for creative writing with an rnn language model." In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*. pp. 1–2.

Rogers, E.M., A. Singhal, and M.M. Quinlan. 2014. "Diffusion of innovations." In *An integrated approach to communication theory and research*. Routledge, pp. 432–448.

Runco, M.A., and S. Acar. 2012. "Divergent thinking as an indicator of creative potential." *Creativity research journal* 24:66–75.

Sadasivan, V.S., A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. 2023. "Can AI-generated text be reliably detected?" *arXiv preprint arXiv:2303.11156*, pp. .

See, A., A. Pappu, R. Saxena, A. Yerukola, and C.D. Manning. 2019. "Do massively pretrained language models make better storytellers?" *arXiv arXiv:1909.10705*, pp. .

Shaikh, M., and E. Vaast. 2023. "Algorithmic interactions in open source work." *Information Systems Research* 34:744–765.

Sharma, A., I.W. Lin, A.S. Miner, D.C. Atkins, and T. Althoff. 2023. "Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support." *Nature Machine Intelligence* 5:46–57.

Sharples, M. 2002. *How we write: Writing as creative design*. Routledge.

Shen, Z., W. Jiang, and Z. Zheng. 2025. "Irrationality-Aware Human Machine Collaboration: Mitigating Alterfactual Irrationality in Copy Trading." *Information Systems Research*, pp. , Ahead of Print.

Sieck, W.R., and H.R. Arkes. 2005. "The recalcitrance of overconfidence and its contribution to decision aid neglect." *Journal of Behavioral Decision Making* 18:29–53.

Siegert, I., R. Böck, and A. Wendemuth. 2014. "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements." *Journal on Multimodal User Interfaces* 8:17–28.

Solaiman, I., M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J.W. Kim, S. Kreps, et al. 2019. "Release strategies and the social impacts of language models." *arXiv preprint arXiv:1908.09203*, pp. .

Stokel-Walker, C. 2022. "AI bot ChatGPT writes smart essays-should academics worry?" *Nature*, pp. .

Storch, N. 2002. "Patterns of interaction in ESL pair work." *Language learning* 52:119–158.

Summers-Stay, D., C.R. Voss, and S.M. Lukin. 2023. "Brainstorm, then select: a generative language model improves its creativity score." In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Sun, L., Y. Yuan, Y. Yao, Y. Li, H. Zhang, X. Xie, X. Wang, F. Luo, and D. Stillwell. 2024. "Large Language Models show both individual and collective creativity comparable to humans." *arXiv preprint arXiv:2412.03151*, pp. .

—. 2025. “Large Language Models show both individual and collective creativity comparable to humans.” *Thinking Skills and Creativity*, pp. 101870.

Susarla, A., R. Gopal, J.B. Thatcher, and S. Sarker. 2023. “The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems.” *Information Systems Research* 34:399–408.

Susnjak, T., and T.R. McIntosh. 2024. “ChatGPT: The end of online exam integrity?” *Education Sciences* 14:656.

Tian, E., and A. Cui. 2023. “GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods.”

Turel, O., and S. Kalhan. 2023. “Prejudiced against the Machine? Implicit Associations and the Transience of Algorithm Aversion.” *Mis Quarterly* 47.

Tweedie, F.J., and R.H. Baayen. 1998. “How variable may a constant be? Measures of lexical richness in perspective.” *Computers and the Humanities* 32:323–352.

Ueoka, H., Y. Murawaki, and S. Kurohashi. 2021. “Frustratingly easy edit-based linguistic steganography with a masked language model.” *arXiv preprint arXiv:2104.09833*, pp. .

Vaccaro, M., A. Almaatouq, and T. Malone. 2024. “When combinations of humans and AI are useful: A systematic review and meta-analysis.” *Nature Human Behaviour*, pp. 1–11.

Van Hout, R., and A. Vermeer. 2007. “Comparing measures of lexical richness.” *Modelling and assessing vocabulary knowledge* 93:115.

Venkatraman, S., A. Uchendu, and D. Lee. 2024. “GPT-who: An information density-based machine-generated text detector.” *Findings of the Association for Computational Linguistics: NAACL*, pp. .

Verspoor, K. 2024. ““Fighting fire with fire”—using LLMs to combat LLM hallucinations.” *Nature*, pp. .

Wallach, H.M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. “Evaluation methods for topic models.” In *Proceedings of the 26th annual international conference on machine learning*. pp. 1105–1112.

Wan, Q., S. Hu, Y. Zhang, P. Wang, B. Wen, and Z. Lu. 2024. “”It Felt Like Having a Second Mind”: Investigating Human-AI Co-creativity in Prewriting with Large Language Models.” *Proceedings of the ACM on Human-Computer Interaction* 8:1–26.

Wang, L., M.I. Mujib, J. Williams, G. Demiris, and J. Huh-Yoo. 2021. “An evaluation of generative pre-training model-based therapy chatbot for caregivers.” *arXiv preprint arXiv:2107.13115*, pp. .

Wang, W., M. Yang, and T. Sun. 2023. “Human-AI Co-Creation in Product Ideation: the Dual View of Quality and Diversity.” *Available at SSRN 4668241*, pp. .

Welleck, S., I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. 2019. “Neural text generation with unlikelihood training.” *International Conference on Learning Representations (ICLR).*, pp. .

Whitecotton, S.M. 1996. “The effects of experience and confidence on decision aid reliance: A causal model.” *Behavioral Research in Accounting* 8:194–216.

Wooldridge, J.M. 2003. “Cluster-sample methods in applied econometrics.” *American Economic Review* 93:133–138.

Wu, T., M. Terry, and C.J. Cai. 2022. “Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts.” In *Proceedings of the 2022 CHI conference on human factors in computing systems*. pp. 1–22.

Wu, Z., D. Ji, K. Yu, X. Zeng, D. Wu, and M. Shidujaman. 2021. “AI creativity and the human-AI co-creation model.” In *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*. Springer, pp. 171–190.

Xu, Y., and D. Reitter. 2018. “Information density converges in dialogue: Towards an information-theoretic model.” *Cognition* 170:147–163.

Yang, D., Y. Zhou, Z. Zhang, T.J.J. Li, and R. LC. 2022. “AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing.” In *Joint Proceedings of the ACM IUI Workshops*. CEUR-WS Team, vol. 10, pp. 1–11.

Yang, J., H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu. 2023. “Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond.” *ACM Transactions on Knowledge Discovery from Data*, pp. .

Yeadon, W., O.O. Inyang, A. Mizouri, A. Peach, and C.P. Testrow. 2023. “The death of the short-form physics essay in the coming AI revolution.” *Physics Education* 58:035027.

Yin, Y., N. Jia, and C.J. Wakslak. 2024. “AI can help people feel heard, but an AI label diminishes this impact.” *Proceedings of the National Academy of Sciences* 121:e2319112121.

Yuan, A., A. Coenen, E. Reif, and D. Ippolito. 2022. “Wordcraft: story writing with large language models.” In *27th International Conference on Intelligent User Interfaces*. pp. 841–852.

Zamfirescu-Pereira, J.D., R.Y. Wong, B. Hartmann, and Q. Yang. 2023. “Why Johnny can’t prompt: how non-AI experts try (and fail) to design LLM prompts.” In *Proceedings of the 2023 CHI conference on human factors in computing systems*. pp. 1–21.

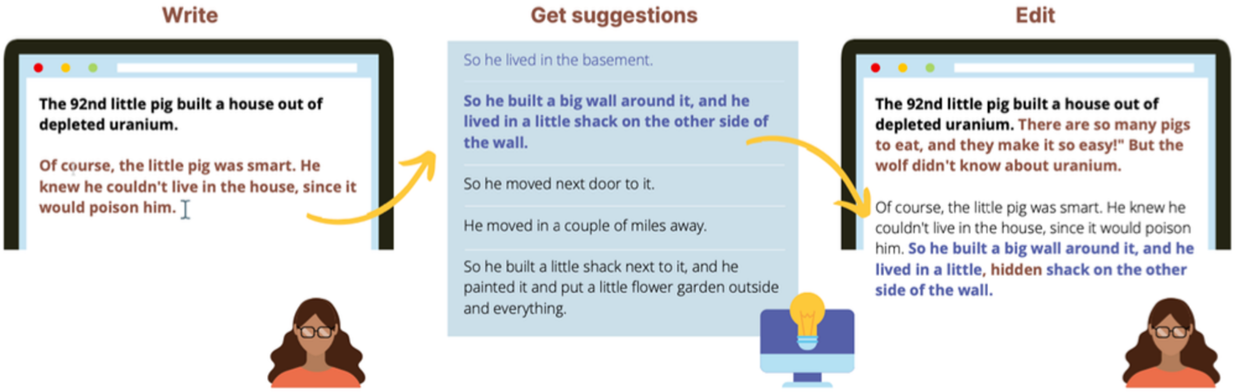
Zanzotto, F.M. 2019. “Human-in-the-loop artificial intelligence.” *Journal of Artificial Intelligence Research* 64:243–252.

Zhang, C., C. Zhang, C. Li, Y. Qiao, S. Zheng, S.K. Dam, M. Zhang, J.U. Kim, S.T. Kim, J. Choi, et al. 2023. “One small step for generative AI, one giant leap for AGI: A complete survey on ChatGPT in AIGC era.” *arXiv preprint arXiv:2304.06488*, pp. .

Zhang, Y., and R. Gosline. 2023. “Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation.” *Judgment and Decision Making* 18:e41.

Zhou, E., and D. Lee. 2024. “Generative artificial intelligence, human creativity, and art.” *PNAS Nexus* 3:page052.

APPENDIX



The 92nd little pig built a house out of depleted uranium.

Of course, the little pig was smart. He knew he couldn't live in the house, since it would poison him.

So he lived in the basement.

So he built a big wall around it, and he lived in a little shack on the other side of the wall.

So he moved next door to it.

He moved in a couple of miles away.

So he built a little shack next to it, and he painted it and put a little flower garden outside and everything.

The 92nd little pig built a house out of depleted uranium. There are so many pigs to eat, and they make it so easy!" But the wolf didn't know about uranium.

Of course, the little pig was smart. He knew he couldn't live in the house, since it would poison him. So he built a big wall around it, and he lived in a little, hidden shack on the other side of the wall.

Figure A1 CoAUTHOR (Lee, Liang, and Yang, 2022), a dataset designed for revealing GPT-3's generative capabilities for interactive writing. Each session starts with a prompt (black text). Writers then freely write (brown), request suggestions from GPT-3 (blue), accept or dismiss suggestions, and edit accepted suggestions or previous texts in any order they choose.

Table A1 10 prompts were retrieved from the WritingPrompts subreddit and used with minor modifications.

Prompt code	Prompt text (Source URL)
shapeshifter	A woman has been dating guy after guy, but it never seems to work out. She's unaware that she's actually been dating the same guy over and over; a shapeshifter who's fallen for her, and is certain he's going to get it right this time. (https://www.reddit.com/r/WritingPrompts/comments/7xihva/wp_a_woman_has_been_dating_guy_after_guy_but_it/)
reincarnation	When you die, you appear in a cinema with a number of other people who look like you. You find out that they are your previous reincarnations, and soon you all begin watching your next life on the big screen. (https://www.reddit.com/r/WritingPrompts/comments/7ezd5t/wp_when_you_die_you_appear_in_a_cinema_with_a/)
mana	Humans once wielded formidable magical power. But with over 7 billion of us on the planet now, Mana has spread far too thinly to have any effect. When hostile aliens reduce humanity to a mere fraction, the survivors discover an old power has begun to reawaken once again. (https://www.reddit.com/r/WritingPrompts/comments/7i3bs6/wp_humans_once_wielded_formidable_magical_power/)
obama	You're Barack Obama. 4 years into your retirement, you awake to find a letter with no return address on your bedside table. It reads "I hope you've had a chance to relax Barack... but pack your bags and call the number below. It's time to start the real job." Signed simply, "JFK." (https://www.reddit.com/r/WritingPrompts/comments/6b3rmg/wp_youre_barack_obama_4_months_into_your/)
pig	Once upon a time there was an old mother pig who had one hundred little pigs and not enough food to feed them. So when they were old enough, she sent them out into the world to seek their fortunes. You know the story about the first three little pigs. This is a story about the 92nd little pig. The 92nd little pig built a house out of depleted uranium. And the wolf was like, "dude." (https://www.reddit.com/r/WritingPrompts/comments/hytfcd/wp_then_the_92nd_little_pig_built_a_house_out_of/)
mattdamon	An alien has kidnapped Matt Damon, not knowing what lengths humanity goes through to retrieve him whenever he goes missing. (https://www.reddit.com/r/WritingPrompts/comments/8p3ora/wp_an_alien_has_kidnapped_matt_damon_not_knowing/)
sideefect	When you're 28, science discovers a drug that stops all effects of aging, creating immortality. Your government decides to give the drug to all citizens under 26, but you and the rest of the "Lost Generations" are deemed too high-risk. When you're 85, the side effects are finally discovered. (https://www.reddit.com/r/WritingPrompts/comments/8on59a/wp_when_youre_28_science_discovers_a_drug_that/)
bee	Your entire life, you've been told you're deathly allergic to bees. You've always had people protecting you from them, be it your mother or a hired hand. Today, one slips through and lands on your shoulder. You hear a tiny voice say "Your Majesty, what are your orders?" (https://www.reddit.com/r/WritingPrompts/comments/88p6rp/wp_your_entire_life_youve_been_told_youre_deathly/)
dad	All of the "#1 Dad" mugs in the world change to show the actual ranking of Dads suddenly. (https://www.reddit.com/r/WritingPrompts/comments/6gl289/wp_all_of_the_1_dad_mugs_in_the_world_change_to/)
isolation	Following World War III, all the nations of the world agreed to 50 years of strict isolation from one another in order to prevent additional conflicts. 50 years later, the United States comes out of exile, only to learn that no one else went into isolation. (https://www.reddit.com/r/WritingPrompts/comments/585ru9/wp_following_world_war_iii_all_the_nations_of_the/)