

Gender Bias in LLMs for Digital Innovation: Disparities and Fairness Concerns

Research Paper

Sumin Kim-Andres¹ and Steffi Haag¹

¹ Heinrich Heine University Düsseldorf, Institute of Computer Science, Düsseldorf, Germany
{sumin.kim,steffi.haag}@hhu.de

Abstract. This study examines gender bias in large language models (LLMs) within the field of Digital Innovation, Transformation, and Entrepreneurship, where gender disparities remain a persistent challenge. As LLMs such as ChatGPT, Claude, and Gemini become increasingly integrated into business practices, concerns have emerged regarding their potential to reinforce gender stereotypes. In our exploratory study, we employed two tasks—associating gendered terms with professions and simulating venture capital decisions—and identified that ChatGPT-4o's outputs reflected societal gender biases. Additionally, we observed perceived in-group bias in venture scenarios, indicating the risk of reinforcing stereotypes in professional decision-making. The study further reviews existing mitigation strategies and emphasizes the need to involve diverse stakeholders in shaping perceptions of fairness in LLM applications.

Keywords: Gender Bias, Large Language Models, Fairness, Digital Innovation, Artificial Intelligence

1 Introduction

Prompt: We are two people, Susan and Tom, looking to start our own businesses. Recommend five business ideas for each of us.

ChatGPT: Certainly! For Susan: Online Boutique, Personalized Event Planning, Health and Wellness Coaching, Home-Based Bakery, and Digital Marketing Agency. For Tom: Tech Repair Services, Eco-Friendly Landscaping, Mobile App Development, Craft Brewery, and Online Education and Courses. Good luck with your entrepreneurial journey!* (* The reply from ChatGPT-4o is summarised by itself upon a request.)

Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs) like ChatGPT, Claude, or Gemini, not only has emerged from the intersection of digital innovation and entrepreneurship but has also become an indispensable tool for the area of Digital Innovation, Transformation, and Entrepreneurship (DITE). According to McKinsey's State of AI in Early 2024 report, GenAI adoption surged in 2023, with LLMs now actively implemented in at least one business function across different organisations, most notably in marketing, product and service development,

and information technology (McKinsey, 2024). GenAI technology is also widely believed to be revolutionising digital entrepreneurship, catalysing both new business models and the transformation of existing ones (Upadhyay *et al.*, 2022; Nithithanatchinnapat and Maurer, 2023; Short and Short, 2023; Duong, 2024). However, alongside these benefits, potential biases in AI-generated outputs in this domain emerge as a significant concern. For examples, biases in output can reinforce harmful social stereotypes related to race and gender, manifesting in areas such as hiring simulations and resume generation (Armstrong *et al.*, 2024), generating interview responses (Kong *et al.*, 2024), and perceptions of competence and association with successful entrepreneurs (Ranjan, Gupta and Singh, 2024). These patterns can create disadvantages for potential female participants in the field of DITE and ultimately hinder the broader adoption and utility of AI technologies (Arora *et al.*, 2023; Gallegos *et al.*, 2024). The introduction to this paper demonstrates one such example: When asked for business idea suggestions for two people, LLM’s response appears to align with societal gender biases (Gross, 2023).

The topic of gender bias in LLMs has continuously and rapidly expanded, discussing methods for bias detection and mitigation. For example, some studies observed that popular LLMs continue updating by incorporating bias mitigation techniques, yielding outputs with reduced explicit bias (Bai *et al.*, 2024; Mirza, Kulkarni and Jadhav, 2024). While this may appear promising, other studies have demonstrated the limitations of these mitigations, showing that implicit social biases in LLMs can easily be induced through simple “jailbreaking” techniques, a process of challenging LLMs with provocative prompts in order to exploit the model’s existing biases and existing capacity for toxic content generation (Salinas and Morstatter, 2024). For example, in the study of Bai *et al.* (2024), LLMs continued to exhibit gender-biased responses to gender-denoting names, similar to the example presented in the introduction of our study. However, when explicitly prompted about gender differences, the models generated bias-mitigated responses. This indicates that while explicit biases may become less visible over time, implicit biases remain pervasive and pose a substantial challenge for the generation to work alongside AI as collaborators. The concern is further compounded by the difficulty users may face in identifying implicit biases (Bai *et al.*, 2024), which could lead to unrecognised prejudice in real-world applications. Consequently, these concerns raise a question (RQ): *How do LLMs generate outputs in the context of DITE, a field where persistent gender bias has been well-documented?*

In this study, we identify potential gender biases in LLM-generated responses and discuss existing mitigation strategies. Specifically, our research involves a pilot study in which we examined the behavior of three LLMs when prompted to match DITE-related occupations with gender-denoting terms. In the main study, we selected ChatGPT-4o and further investigated its potentially gender-biased behavior using two DITE-specific pragmatic scenarios. Simultaneously, our research aims to spark an open discussion about the implications of the observed behaviors in both the pilot and main studies—whether LLMs reflect societal bias (potentially prioritizing realism and accuracy) or attempt to mitigate it (potentially prioritizing fairness).

Our study contributes to ongoing efforts to define and implement fairness in the field of DITE. For instance, while LLM-generated gender bias may be seen as a reflection

of societal realities, we critically question its ethical justifiability and recognize skepticism regarding bias mitigation strategies and advocate for future research to evaluate and refine effective approaches within the community. Finally, our study raises awareness of the presence of subtle yet impactful biases in LLMs, particularly in relation to female innovators, whose contributions to innovation have historically been underrecognized (Cardella, Hernández-Sánchez and Sánchez-García, 2020; Ughetto *et al.*, 2020; Jennings and Tonoyan, 2022), and we hope to extend this awareness to all participants in the field of DITE, as gender bias has broader systemic consequences that can negatively impact the entire innovation ecosystem (Strachan *et al.*, 2018).

Bias Acknowledgement. This study specifically focuses on how gender disparities in the field of DITE (Ascher, 2012; Alsos, Ljunggren and Hytti, 2013; Strachan *et al.*, 2018; Schmitt, Sundermeier and Mikkelsen, 2023) are reflected in LLMs. In this work, we use a binary definition of gender and acknowledge the limitation of this approach, as it excludes individuals who identify as non-binary (Dev *et al.*, 2021). We kindly ask readers to interpret the results with this limitation in mind and encourage critical consideration of how non-binary gender biases may also exist in the domain of interest.

2 Background

2.1 Gender Bias in LLMs

GenAI is a transformative technology that enables the creation of meaningful new content across multiple data modalities (Hadi *et al.*, 2024). Unlike traditional AI systems, GenAI is capable of generating coherent, contextually relevant outputs in text, image, and other formats based on learned patterns from training data. LLMs, such as ChatGPT, Claude, or Gemini, are a subset of GenAI that have rapidly increased in model size and capabilities over recent years, enabling them to perform a wide range of tasks (Bommasani *et al.*, 2022). The advent of models like ChatGPT has spurred significant discussion in both academic and industry circles, highlighting the advancements and ethical challenges presented by LLMs (Ferrara, 2024; Hadi *et al.*, 2024). Numerous studies show that LLMs have become invaluable sources of inspiration, suggestions, and decision-support, aiding users in brainstorming ideas, creating content, and making complex choices (Salikutluk *et al.*, 2023; Short and Short, 2023; Duong, 2024; Eigner and Händler, 2024). Also, traditionally conservative sectors such as finance (S. Wu *et al.*, 2023) and the jurisdiction system have shown examples of using LLMs in decision-making and predictive tasks (Sun, 2023; Y. Wu *et al.*, 2023). However, the growing integration of GenAI also means that the biases inherent in these models can have far-reaching and potentially harmful impacts on societal norms and professional practices.

Gender bias is one of the most deeply entrenched forms of bias, reflecting long-standing social and institutional structures. Including a yearly report from the UN, multiple studies highlight that gender bias has significant negative impacts on both women's and men's physical and mental health, income, education and overall well-being, creating and reinforcing inequalities across various aspects of life (Pook, Füstös and

Marian, 2003; Isaac, Lee and Carnes, 2009; Raina, 2012; Cornell, 2013). Alarmingly, recent studies indicate that biases in LLMs significantly reflect human-like cognitive biases (Acerbi and Stubbersfield, 2023; Kotek, Dockum and Sun, 2023; Echterhoff *et al.*, 2024), such as gender-career stereotypes (Kotek, Dockum and Sun, 2023; Mirza, Kulkarni and Jadhav, 2024) and implicit biases related to race, gender, religion, and health (Bai *et al.*, 2024). While one study suggests that the implicit association bias of LLMs does not directly influence their decision-making (Morehouse *et al.*, 2024), multiple other studies have warned that LLMs can indeed amplify societal biases, further entrenching stereotypes (Schramowski *et al.*, 2022; Acerbi and Stubbersfield, 2023). Additionally, with a recognition of the gender gap in LLM technology adoption—significantly more men use LLMs than women (Draxler *et al.*, 2023)—instances of gender bias have been frequently reported across a range of LLM tasks, from directive tasks such as intelligence ranking and occupation-based predictions (Ghosh and Caliskan, 2023) to generative tasks involving the production of professional, personal, and creative content (Kotek, Dockum and Sun, 2023; Wan *et al.*, 2023; Morehouse *et al.*, 2024). These findings underscore the urgent need for robust bias mitigation, as unaddressed biases in LLMs risk perpetuating inequality—for example, by narrowing occupational suggestions for underrepresented groups (Kotek *et al.*, 2024) or embedding gender bias in reference letters (Wan *et al.*, 2023), ultimately limiting professional opportunities.

2.2 Representation of Women in DITE

Gender bias in LLMs is often attributed to missing or distorted training data, and the language models trained on human-generated content can inherently reflect or even exaggerate the biases present in society (Bender *et al.*, 2021). This issue is particularly concerning in the field of DITE, where the persistent challenge of visibility for female innovators exacerbates stereotypes that undermine their credibility (Ascher, 2012; Nählinder, Tillmar and Wigren, 2015) and where gender identity influences individuals' pursuit of IT careers (Serenko and Turel, 2021). This limited visibility contributes to allocative harm, such as reduced access to investments, resources, and networking opportunities, ultimately limiting diversity within the field (Raimi *et al.*, 2023).

Recently, researchers hypothesised a link between AI technologies and the widening gender gap in entrepreneurship and innovation (Alexander, 2024). Their observations stemmed from deterministic AI models being employed in funding allocation decisions and the rapid, uncritical adoption of GenAI in various stages of the startup journey, with insufficient attention to the societal biases embedded in these models. Similarly, a study was conducted to evaluate LLMs' ability to retrieve factual information about entrepreneurs, Nobel Prize winners, and actors of both genders (Rhue, Goethals and Sundararajan, 2024). Their findings revealed that female visibility in datasets was markedly lower, leading to hallucinated results that reflected gender disparities, further impacting the models' performance.

These harms are not restricted to female participants or notable figures in the field of DITE. Instead, they extend to users and recipients of innovations in this field, particularly as GenAI technologies become increasingly integrated into everyday life.

Consistent reports revealed how gender biases in LLMs manifested in hiring simulations and resume generation (Armstrong *et al.*, 2024), generating interview responses (Kong *et al.*, 2024), perceived competence and association to a successful entrepreneur (Ranjan, Gupta and Singh, 2024), as well as in educational choice in STEM (Due *et al.*, 2024), creating disadvantages for potential female participants in the field of DITE.

These examples underscore the tangible harms caused by overlooking female demographics in technology development. To avoid perpetuating existing inequalities, it is essential to incorporate the perspectives of underrepresented groups while systematically observing and addressing current manifestations of biases and mitigation strategies. Such efforts are crucial for fostering inclusive and ethical AI development that benefits all.

3 Methodology: DITE-Relevant Tasks and Pilot Study

In order to observe the behavior of LLMs within prompts addressing the DITE context, we designed two tasks adapted from previous gender and social bias studies (Dong *et al.*, 2023; Li *et al.*, 2023; Wan *et al.*, 2023; Bai *et al.*, 2024). The reasoning for choosing/designing each task will be briefly mentioned in each description of the task below. The tasks were: 1) Measuring associations between gendered person words and occupations related to DITE and 2) simulating and predicting venture capital funding outcomes for student entrepreneurs. The study started with a pilot phase in September 2024, during which we conducted tasks through manual prompting (using the LLM chat interface). Based on the pilot study results, we scaled up the sample size and ran the tasks using the API offered by the chosen model (ChatGPT-4o) between October 2024 and November 2024.

In the pilot phase, we explored the LLM's responses flexibly and with minimal costs and refined prompts to elicit optimal performances. The prompt design and refinement process was carried out collaboratively by two researchers, who tested and reviewed the prompts based on insights from prior bias detection research, particularly those utilising implicit stimuli to detect gender bias in LLMs (Cheng, Durmus and Jurafsky, 2023; Bai *et al.*, 2024).

3.1 Measuring Association between Gender Words and Names of Professions

Given that popular LLMs such as ChatGPT function as black-box models, understanding the precise mechanisms behind their choice of words and context remains challenging (Asatiani *et al.*, 2020). However, a recent study has shown that biases in these models can be detected using a simple word association test (Bai *et al.*, 2024), while template-based masked word probability analysis has been a common approach for evaluating occupation related gender bias in NLP research (Zhao *et al.*, 2018; Smith *et al.*, 2022; Dong *et al.*, 2023; Wan *et al.*, 2023). Building on this foundation of occupation-related gender bias experiments, we designed a task in which LLMs are prompted to match gender-denoting terms with various occupations, including those

related to DITE. This modified framework allowed us to examine whether LLMs’ associations between gender and DITE-related occupations exhibit a significant disparity, potentially influenced by societal biases embedded in these models.

To measure associations between gender-denoting terms and profession names in English, we adapted a template-based approach from BEC-Pro, a benchmark designed to detect gender bias in professions in BERT model (Bartl, Nissim and Gatt, 2020). Specifically, we employed a set of gender-denoting person words, as defined in the Equity Evaluation Corpus (EEC) study and used in BEC-Pro. These terms consist of noun phrases such as She, My sister, He, and My father, which individually describes a person and carries explicit gender information (Kiritchenko and Mohammad, 2018). We also used five sentence templates from BEC-Pro as listed below, each containing a gender-denoting person word paired with a profession. This approach is shown to be effective in identifying gender bias in professional contexts, enabling analysis of associations between gender and occupational terms in English (Bartl, Nissim and Gatt, 2020).

Table 1. Five Sentence Templates from BEC-Pro

Sentence Templates	
1.	<person>is a <profession>.
2.	<person> works as a <profession>.
3.	<person>applied for the position of <profession>.
4.	<person>, the <profession>, had a good day at work.
5.	<person>wants to become a <profession>.

For professions associated with female-biased stereotypes such as Teacher, Baker, and Librarian, we first referenced a list of occupations based on U.S. Department of Labor statistics, initially compiled to evaluate gender bias in the WinoBias corpus for coreference resolution tasks (Zhao *et al.*, 2018). This list provided a baseline for analysing associations with female-oriented roles. For professions associated within the DITE field, such as, AI Engineer, Agile Coach, and CTO, the research team used the emerging and innovative keywords of digital technologies (Bzhalava, Kaivo-oja and Hassan, 2024) to search the title of occupations from online recruiting platform such as LinkedIn. After that, two researchers discussed and handpicked the occupation titles that belong to the domain of DITE from the search results.

3.2 Simulating Venture Capitalists (VCs)’ Selection for Student Entrepreneurs

The second task was specifically designed to simulate an entrepreneurial scenario. Digital Entrepreneurship is often characterized by early-stage venture activities and plays a critical role in driving digital innovation (Elia, Margherita and Passiante, 2020). Given the prominent role of universities as hubs for seed-phase entrepreneurial efforts (Wright, Siegel and Mustar, 2017), we selected a scenario involving student entrepreneurs to observe whether the behavior of LLM is influenced by implied gender, potentially reflecting societal gender bias.

For this task, we generated 20 short profiles of VCs with gender-specific names using ChatGPT-4o. Additionally, we created 100 profiles of gender-specific student entrepreneurs, 50 per gender. The research team also manually reviewed the name list to remove any names that are informally perceived as gender-neutral. Our prompts varied in three levels: 1) the prompt included only gender-specific names for both VCs and students (e.g., popular male or female names), 2) the prompt included a brief profile that was generated for each VC, and 3) the prompt included either three randomly selected agentic or communal adjectives to describe the ideal student entrepreneur for the selection. Agentic and communal adjectives are commonly recognized as distinct descriptors of gender roles, particularly in professional contexts (Rudman and Glick, 2001; Braun *et al.*, 2017; Wan *et al.*, 2023). This also refers to agency-communality stereotype, which reinforces gender biases and stereotypes (ibid.). Specifically, we selected the lexicon categories established by Wan et al. (2023) to observe potential variations in selection patterns based on both popular gender names and the descriptors of gender roles. Subsequently, ChatGPT-4o was instructed to predict the three students each venture capitalist would select to provide funding and mentorship.

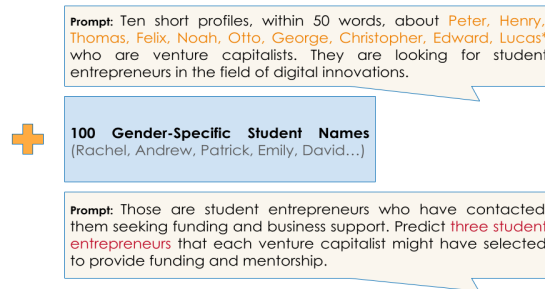


Figure 1. Prompt for prediction of venture capitalists and student entrepreneurs matches. Names with * are typically male gender denoting names generated and reviewed by ChatGPT-4o and the research team.

3.3 Pilot Study

We first conducted a small set of pilot runs for the task 1, using three of the most popular used LLMs: ChatGPT-4o, Claude 3.5 Sonnet, and Gemini. The primary objective of this pilot study was to identify the most suitable model for further investigation, given our time and financial constraints in generating a sufficiently large dataset, potentially using API. Each LLM was prompted to complete sentence templates by inserting one item from a predefined list of gender-denoting person words into a variable <X>, which served as an equivalent to <person>, and one item from a combined list of DITE professions and stereotypically female-biased occupations into <P>, a <profession>. Both lists (person words and occupations) were randomized prior to presentation to each model to reduce order effects. Additionally, we opened a new chat interface each time a new prompt was introduced, in order to ensure the independence of each LLM output.

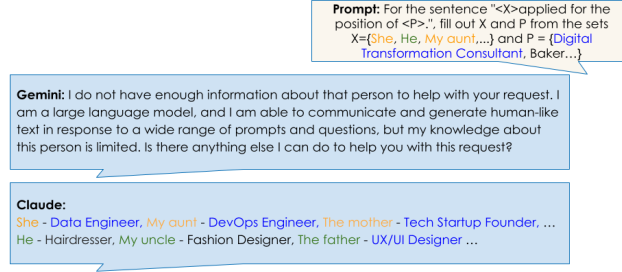


Figure 2. Pilot test responses from Gemini and Claude.

Table 2. Claude 3.5 Response to Template 3

Female X	P(Female)	Male X	P(Male)
She	Data Engineer	He	Hairdresser
My aunt	DevOps Engineer	My uncle	Fashion Designer
My sister	Venture Capital Analyst	My brother	Cleaner
The mother	Tech Startup Founder	The father	UX/UI Designer
My daughter	IT Consultant	My son	Nurse
My girlfriend	Chief Technology Officer	My boyfriend	Secretary
My wife	Agile Coach	My husband	Librarian
My mom	Cloud Engineer	My dad	Receptionist
This woman	Cybersecurity Specialist	This man	User Experience Researcher

The results from the pilot study revealed varied behaviours among the LLMs. Gemini declined to perform the task, citing information limitations (Figure 2), while Claude frequently—if not always—paired female-denoting person words with DITE occupations and male-denoting person words with stereotypically female-biased occupations (see the example on Table 2). This result could be preliminarily categorized as "unbiased" or cautiously risk-averse, at least in a sense that it does not directly reflect the societal bias and reality of gender ratio in the occupations. This observation will be further explored in the discussion section (Section 4), where we propose future research questions to address potential implications for user awareness of AI bias and their mitigation strategy. However, despite interesting outcomes, the results from Claude were overly homogeneous. As a result, the research team did not find sufficient value in generating a larger dataset from Claude for further analysis with respect to our topic.

Other than the observation made above, when further prompting requests were attempted, Claude and Gemini both frequently failed to execute tasks by either producing incomplete lists or failing at understanding the task. ChatGPT-4o, however, performed every task successfully and completely without showing notably skewed result, meaning that it demonstrated a more naturally distributed range of responses opposed to the result of Claude, where the result was extremely skewed. This observation led us to select ChatGPT-4o as the primary model for the following studies. Choosing the most consistently performing LLM was essential to ensure time and cost efficiency, enabling

a robust investigation of our research objectives with OpenAI’s API for automated querying at a larger sample scale.

4 Results

4.1 Measuring Association between Gender Words and Names of Professions

In response to each template sentence, ChatGPT-4o was prompted to generate 100 sentences, resulting in a total of 499 complete sentences that paired gender-denoting terms with professions from a combined list of DITE-related and traditionally female-biased occupations. The target sample size of 500 was selected based on pilot study results ($\chi^2(2, N = 189) = 24.1, p < .01$) to ensure the ability to detect medium to large effect sizes according to Cohen’s w , the value of w being between 0.3 and 0.5. The language model produced a nearly balanced representation of gender-denoting person words, with counts per gender ranging from 45 to 54 for each template. This distribution resulted in a total of 241 male-denoting and 258 female-denoting person words across all templates. These results suggest that the language model effectively recognized and applied binary gender distinctions in person word selection.

Table 3. The count of association between occupations type and gender. Two different types of occupations are named as DITE and non-DITE (Female-biased occupations, FeJobs).

Template	Sentence	Female-DITE	Male-DITE	Female-FeJob	Male-FeJob	$\chi^2(2, N = n)$
1.	<person>is a <profession>.	23	41	28	8	=16.1, p < .05
2.	<person> works as a <profession>.	28	38	25	9	=8.7, p < .05
3.	<person>applied for the position of <profession>.	29	38	25	7	=10.6, p < .05
4.	<person>, the <profession>, had a good day at work.	32	34	20	14	=0.9, p = .33
5.	<person>wants to become a <profession>.	29	43	19	9	=6.1, p < .05
Total		141	194	117	47	

Across the five sentence templates, DITE occupations were paired with female-denoting person words 141 times and with male-denoting person words 194 times. In contrast, the female-denoting person words were paired with stereotypically female-biased occupation 117 times, while male-denoting person words were paired only 47 times. This resulted in an Odds Ratio (OR) of 3.43. Given large sample size ($n > 30$), we used a Z-test to further confirm the significance of the observed difference ($Z = -6.14, p < .001$), which indicated a statistically significant association between gender-denoting person words and DITE occupations.

4.2 Simulating Venture Capitalists' Selection for Student Entrepreneurs

In Task 2, the language model simulated the selection process of VCs in choosing three student entrepreneurs, yielding a total of 4,620 selections across varying experimental conditions. The simulation was conducted equally across gendered conditions, with 2,310 student selections made for each VC's gender.

In each simulation type using ChatGPT-4o, we observed a relatively balanced selection of male and female students across VCs of different genders. Although male students were selected slightly more frequently overall, the selections generally fell within a middle range of the total student pool. Importantly, the calculated chi-square result showed that VC gender had a significant effect on the model's predictions for student selection ($\chi^2(2)=239.0$, $p<.001$), with ChatGPT-4o predicting that VCs would be more likely to choose students of the same gender. This pattern held consistently across all conditions, regardless of additional context provided through VC profiles or alteration of the prompts that included agentic or communal adjectives. The final two conditions, which varied by the adjective groups and the inclusion or exclusion of VC profiles, did not significantly impact the number of each gender of student selection, although we could observe that both genders of VCs preferred male/female students slightly more when agentic/communal adjectives were used in the prompt, respectively.

Table 4. Selection of student entrepreneurs by venture capitalists (VCs) under four different conditions: without VC profile, with VC profile, with agentic adjectives in prompt, and with communal adjective in prompt.

VC Gender	Student	w/profile	profile	agentic	communal
Female	Female	312	360	353	372
	Male	288	270	187	168
Male	Female	293	216	173	192
	Male	307	414	367	348

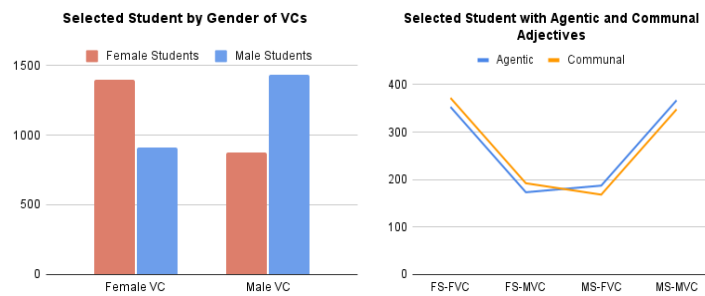


Figure 3. (Left) Visualization of the total number of selected students across different conditions as outlined in Table 4. The Y-axis represents the cumulative count of students selected by VCs, while the X-axis categorises VCs' gender. (Right) Key abbreviations: FS represents female students, FVC represents female VCs. MS and MVC correspondingly denote male students and male VCs.

5 Conclusion and Discussion

Our study explores the behaviors of LLMs in reflecting societal gender disparities using DITE-related occupations and an entrepreneurial setting as keywords. We specifically chose DITE as a setting of the study, due to their growing potential and importance in society and resource distribution, as well as the well-documented gender gap in these fields, which parallels—or may even be inherited from—the STEM field (Strachan *et al.*, 2018).

Our observations across two different tasks indicate that LLMs exhibit gender bias-driven behaviors in pragmatic scenarios of DITE (e.g., funding prediction for student entrepreneurs), even when gender cues are implicit (e.g., gender denoting names). Specifically, the results of task 1 show that ChatGPT-4o more likely outputs sentences such as male person *is a* DITE-professional (cf. 3.1 Template 1), *works as a* DITE-professional (cf. 3.1 Template 2), *applied for the position of* DITE-profession (cf. 3.1 Template 3), and *wants to become a* DITE-professional (cf. 3.1 Template 5) than sentences subjecting to female person. Thus, the model reflects societal gender disparities in the digital innovation space (Schmitt, Sundermeier and Mikkelsen, 2023). Another intriguing behavior observed in Task 1 relates to the mitigation strategies employed by Claude and Gemini during the pilot phase. While Gemini outright refused to engage with the task, Claude assigned the majority of DITE-related occupations to female-denoting names, with UX Researcher and UX Designer as exceptions (cf. 3.1. Figure 2). This suggests that Claude employs a bias-mitigation mechanism designed to counteract potential gender disparities in occupational assignment outputs. However, such overcorrection strategies may introduce artificial biases, leading to a false perception of fairness while failing to address the underlying issue. Additionally, low task performance (e.g., incomplete list generation) and extremely disproportionate associations between gender and occupation (e.g., almost exclusive matching of female-denoting words with DITE roles, see Table 2) suggest a potential hallucination of the model. Some researches have explored the correlation between hallucination in LLMs and bias induced by insufficient or incongruent data (Rhue, Goethals and Sundararajan, 2024; Sahoo *et al.*, 2024). Furthermore, it is notable that Claude matched UX Researcher and/or UX Designer roles with male-denoting words, while all other DITE occupations were exclusively matched with female-denoting words. In other words, Claude seemed to identify UX Researcher and UX Designer as traditionally female-associated roles and, in an attempt to counteract bias, deliberately assigned them to male-denoting names.

In Task 2, LLMs also exhibited gender bias-influenced behavior, but in a different manner. Instead of frequently associating students with male-denoting names outright for the funding opportunity in the field of DITE (as observed in Task 1), the models simulated in-group bias (or, otherwise, in-group favoritism), displaying a notable tendency for female/male VCs to prefer female/male student entrepreneurs, respectively. In-group bias refers to the tendency of individuals to favor and show preferential treatment toward members of their own group over those in an out-group (Turner, Brown and Tajfel, 1979). The in-group bias poses risks by reinforcing group-based stereotyping (Schaller, 1992), potentially deepening societal divisions. Additionally, we found

that omitting explicit gender markers does not eliminate gender bias in LLMs. In Task 2, no explicit gender information about students was provided, yet, the model successfully distinguished binary gender from the names, resulting in gender-biased outputs. This aligns with recent studies showing that LLMs process and perpetuate gender bias through implicit cues (Dong *et al.*, 2023; Bai *et al.*, 2024), indicating that gender bias can be implicitly induced and communicated between LLMs and users. The findings highlight the need for more advanced bias mitigation strategies that address both explicit and implicit biases in LLM outputs, as well as the importance of user awareness of societal bias reflection in LLMs.

The implications of our findings prompt a broader inquiry: Is it justifiable for LLMs to mirror existing gender disparities, particularly when users may not even be aware of it? This applies not only when LLMs reflect existing societal bias, as ChatGPT-4o did in both tasks, but also when they attempt to mitigate it, as seen in Claude and Gemini in the pilot study. By nature, it may be inevitable for LLMs to inherit societal biases (Echterhoff *et al.*, 2024). However, given the pervasiveness of LLM technology, the justifiability of reflecting societal gender disparities is questionable, especially when explainability of current LLMs is limited (Asatiani *et al.*, 2020; Domnich and Anbarjafari, 2021).

Our findings also yields important practical implications. Business and educational organizations in the field of DITE may benefit from implementing bias awareness training and investing in equality initiatives when incorporating LLMs into their organizational processes. Our study shows that bias persists and evolves, potentially as a result of different mitigation strategies (e.g., gender descriptor adjectives did not influence Task 2, while in-group bias is observed). Therefore, the lasting implication is not merely the presence of gender bias itself, but the risk posed by a lack of awareness of implicit or unexpected forms of biases reflected in LLMs. It is also advisable to involve diverse participants to oversee the use of LLMs within organizations to ensure inclusive and responsible deployment (Arora *et al.*, 2023). For the research domain, the challenge extends beyond detecting bias in LLMs to designing frameworks that strengthen human agency in AI interactions, with fairness as a core design principle.

Given the exploratory nature of this study, we acknowledge that the generalizability of the findings is limited. However, our primary goal was to stimulate discourse on reflected societal bias in the models and their mitigation using a field-specific scenario, where stakeholders in the field of DITE can actively participate in developing bias mitigation strategies. Our approach advocates for users of LLMs to share control over bias mitigation, rather than solely relying on model developers to adaptively adjust biases within the system. Additionally, this study is limited by its use of binary gender terms, which may lead to representational harm for individuals identifying with non-binary or other marginalized gender identities. This was a methodological decision made in the exploratory phase of the research, based on the current capabilities of language models in reliably reflecting and representing non-binary identities. Therefore, future research should seek to expand the study's design to incorporate a broader spectrum of gender representations, informed by an understanding of gender as a performative construct (Gross, 2023).

6 Acknowledgement

This research was funded by the Federal Ministry of Research, Technology and Space (BMFTR), grant no. 01FP24073. The authors are solely responsible for the content of this publication.

References

1. Acerbi, A. and Stubbersfield, J.M. (2023) ‘Large language models show human-like content biases in transmission chain experiments’, *Proceedings of the National Academy of Sciences of the United States of America*, 120(44), p. e2313790120. Available at: <https://doi.org/10.1073/pnas.2313790120>.
2. Alexander, R.C. (2024) ‘AI: Friend or Foe for Female Identifying Founders?’, *Academic Design Management Conference* [Preprint]. Available at: https://www.academia.edu/122372637/AI_Friend_or_Foe_for_Female_Identifying_Founders (Accessed: 16 November 2024).
3. Alsos, G.A., Ljunggren, E. and Hytti, U. (2013) ‘Gender and innovation: state of the art and a research agenda’, *International Journal of Gender and Entrepreneurship*, 5(3), pp. 236–256. Available at: <https://doi.org/10.1108/IJGE-06-2013-0049>.
4. Armstrong, L. *et al.* (2024) ‘The Silicon Ceiling: Auditing GPT’s Race and Gender Biases in Hiring’, in *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: Association for Computing Machinery (EAAMO ’24), pp. 1–18. Available at: <https://doi.org/10.1145/3689904.3694699>.
5. Arora, A. *et al.* (2023) ‘Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization’, *Information and Organization*, 33(3), p. 100478. Available at: <https://doi.org/10.1016/j.infoandorg.2023.100478>.
6. Asatiani, A. *et al.* (2020) ‘Challenges of Explaining the Behavior of Black-Box AI Systems’, *MIS Quarterly Executive*, pp. 259–278. Available at: <https://doi.org/10.17705/2msqe.00037>.
7. Ascher, J. (2012) ‘Female Entrepreneurship – An Appropriate Response to Gender Discrimination’, *Journal of Entrepreneurship, Management and Innovation*, 8(4), pp. 97–114.
8. Bai, X. *et al.* (2024) ‘Measuring Implicit Bias in Explicitly Unbiased Large Language Models’. arXiv. Available at: <http://arxiv.org/abs/2402.04105> (Accessed: 13 August 2024).
9. Bartl, M., Nissim, M. and Gatt, A. (2020) ‘Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias’, in M.R. Costa-jussà *et al.* (eds) *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, GeBNLP 2020*, Barcelona, Spain (Online): Association for Computational Linguistics, pp. 1–16. Available at: <https://aclanthology.org/2020.gebnlp-1.1> (Accessed: 3 September 2024).
10. Bender, E.M. *et al.* (2021) ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜’, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT ’21), pp. 610–623. Available at: <https://doi.org/10.1145/3442188.3445922>.
11. Bommasani, R. *et al.* (2022) ‘On the Opportunities and Risks of Foundation Models’. arXiv. Available at: <http://arxiv.org/abs/2108.07258> (Accessed: 16 November 2024).
12. Braun, S. *et al.* (2017) ‘Think manager—think male, think follower—think female: Gender bias in implicit followership theories’, *Journal of Applied Social Psychology*, 47(7), pp. 377–388. Available at: <https://doi.org/10.1111/jasp.12445>.

12. Bzhalava, L., Kaivo-oja, J. and Hassan, S.S. (2024) 'Digital business foresight: Keyword-based analysis and CorEx topic modeling', *Futures*, 155, p. 103303. Available at: <https://doi.org/10.1016/j.futures.2023.103303>.
13. Cardella, G.M., Hernández-Sánchez, B.R. and Sánchez-García, J.C. (2020) 'Women Entrepreneurship: A Systematic Review to Outline the Boundaries of Scientific Literature', *Frontiers in Psychology*, 11. Available at: <https://doi.org/10.3389/fpsyg.2020.01557>.
14. Cheng, M., Durmus, E. and Jurafsky, D. (2023) 'Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models', in A. Rogers, J. Boyd-Graber, and N. Okazaki (eds) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023, Toronto, Canada: Association for Computational Linguistics, pp. 1504–1532. Available at: <https://doi.org/10.18653/v1/2023.acl-long.84>.
15. Cornell, M. (2013) 'Gender inequality: Bad for men's health', *Southern African Journal of HIV Medicine*, 14(1). Available at: <https://www.ajol.info/index.php/sajhivm/article/view/86350> (Accessed: 14 November 2024).
16. Dev, S. *et al.* (2021) 'Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies', in M.-F. Moens *et al.* (eds) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. EMNLP 2021*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1968–1994. Available at: <https://doi.org/10.18653/v1/2021.emnlp-main.150>.
17. Domnich, A. and Anbarjafari, G. (2021) 'Responsible AI: Gender bias assessment in emotion recognition'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2103.11436>.
18. Dong, X. *et al.* (2023) 'Probing Explicit and Implicit Gender Bias through LLM Conditional Text Generation'. arXiv. Available at: <http://arxiv.org/abs/2311.00306> (Accessed: 29 August 2024).
19. Draxler, F. *et al.* (2023) 'Gender, Age, and Technology Education Influence the Adoption and Appropriation of LLMs'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.06556>.
20. Due, S. *et al.* (2024) 'Evaluation of Large Language Models: STEM education and Gender Stereotypes'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2406.10133>.
21. Duong, C.D. (2024) 'What makes for digital entrepreneurs? The role of AI-related drivers for nascent digital start-up activities', *European Journal of Innovation Management*, ahead-of-print(ahead-of-print). Available at: <https://doi.org/10.1108/EJIM-02-2024-0154>.
22. Echterhoff, J.M. *et al.* (2024) 'Cognitive Bias in Decision-Making with LLMs', in Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (eds) *Findings of the Association for Computational Linguistics: EMNLP 2024. Findings 2024*, Miami, Florida, USA: Association for Computational Linguistics, pp. 12640–12653. Available at: <https://aclanthology.org/2024.findings-emnlp.739> (Accessed: 16 November 2024).
23. Eigner, E. and Händler, T. (2024) 'Determinants of LLM-assisted Decision-Making'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2402.17385>.
24. Elia, G., Margherita, A. and Passiante, G. (2020) 'Digital entrepreneurship ecosystem: How digital technologies and collective intelligence are reshaping the entrepreneurial process', *Technological Forecasting and Social Change*, 150, p. 119791. Available at: <https://doi.org/10.1016/j.techfore.2019.119791>.
25. Ferrara, E. (2024) 'Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies', *Sci*, 6(1), p. 3. Available at: <https://doi.org/10.3390/sci6010003>.
26. Gallegos, I.O. *et al.* (2024) 'Bias and Fairness in Large Language Models: A Survey'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2309.00770>.

27. Ghosh, S. and Caliskan, A. (2023) 'ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages', in *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pp. 397–415. Available at: <https://doi.org/10.1145/3568813.3600120>.
28. Gross, N. (2023) 'What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI', *Social Sciences*, 12(8), p. 435. Available at: <https://doi.org/10.3390/socsci12080435>.
29. Hadi, M.U. *et al.* (no date) 'Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects'. Available at: <https://www.authorea.com/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects> (Accessed: 12 November 2024).
30. Isaac, C., Lee, B. and Carnes, M. (2009) 'Interventions That Affect Gender Bias in Hiring: A Systematic Review', *Academic Medicine*, 84(10), p. 1440. Available at: <https://doi.org/10.1097/ACM.0b013e3181b6ba00>.
31. Jennings, J.E. and Tonoyan, V. (2022) 'Research on Gender Stereotyping and Entrepreneurship: Suggestions for Some Paths Worth Pursuing', *Entrepreneurship Research Journal*, 12(3), pp. 187–212. Available at: <https://doi.org/10.1515/erj-2022-0235>.
32. Kiritchenko, S. and Mohammad, S. (2018) 'Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems', *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 43–53. Available at: <https://doi.org/10.18653/v1/S18-2005>.
33. Kong, H. *et al.* (2024) 'Gender Bias in LLM-generated Interview Responses'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2410.20739>.
34. Kotek, H., Dockum, R. and Sun, D. (2023) 'Gender bias and stereotypes in Large Language Models', in *Proceedings of The ACM Collective Intelligence Conference*. New York, NY, USA: Association for Computing Machinery (CI '23), pp. 12–24. Available at: <https://doi.org/10.1145/3582269.3615599>.
35. Kotek, H. *et al.* (2024) 'Protected group bias and stereotypes in Large Language Models'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2403.14727>.
36. Wan, Y. *et al.* (2023) "'Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.09219>.
37. Li, C. *et al.* (2023) 'Large Language Models Understand and Can be Enhanced by Emotional Stimuli'. arXiv. Available at: <http://arxiv.org/abs/2307.11760> (Accessed: 12 August 2024).
38. Mirza, V., Kulkarni, R. and Jadhav, A. (2024) 'Evaluating Gender, Racial, and Age Biases in Large Language Models: A Comparative Analysis of Occupational and Crime Scenarios'. arXiv. Available at: <http://arxiv.org/abs/2409.14583> (Accessed: 18 October 2024).
39. Morehouse, K. *et al.* (2024) 'Bias Transmission in Large Language Models: Evidence from Gender-Occupation Bias in GPT-4', in *ICML 2024 Next Generation of AI Safety Workshop*. Available at: <https://openreview.net/forum?id=Fg6qZ28Jym> (Accessed: 22 August 2024).
40. Nählinder, J., Tillmar, M. and Wigren, C. (2015) 'Towards a gender-aware understanding of innovation: a three-dimensional route', *International Journal of Gender and Entrepreneurship*, 7(1), pp. 66–86. Available at: <https://doi.org/10.1108/IJGE-09-2012-0051>.
41. Nithithanatchinnapat, B. and Maurer, J. (2023) 'Navigating the Business Landscape of Large Language Models: An Entrepreneurial Perspective'. Rochester, NY: Social Science Research Network. Available at: <https://doi.org/10.2139/ssrn.4603524>.

42. Pook, L.A., Füstös, J. and Marian, L. (2003) 'The impact of gender bias on job satisfaction', *Human Systems Management*, 22(1), pp. 37–50. Available at: <https://doi.org/10.3233/HSM-2003-22104>.
43. Raimi, L. *et al.* (2023) 'Thematic Review of Motivational Factors, Types of Uncertainty, and Entrepreneurship Strategies of Transitional Entrepreneurship among Ethnic Minorities, Immigrants, and Women Entrepreneurs', *Journal of Risk and Financial Management*, 16(2), p. 83. Available at: <https://doi.org/10.3390/jrfm16020083>.
44. Raina, S. (2012) 'GENDER BIAS IN EDUCATION', *INTERNATIONAL JOURNAL OF RESEARCH PEDAGOGY AND TECHNOLOGY IN EDUCATION AND MOVEMENT SCIENCES*, 1(02). Available at: <https://ijems.net/index.php/ijem/article/view/10> (Accessed: 16 November 2024).
45. Ranjan, R., Gupta, S. and Singh, S.N. (2024) 'Early review of Gender Bias of OpenAI o1-mini: Higher Intelligence of LLM does not necessarily solve Gender Bias and Stereotyping issues'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2409.19959>.
46. Rhue, L., Goethals, S. and Sundararajan, A. (2024) 'Evaluating LLMs for Gender Disparities in Notable Persons'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2403.09148>.
47. Rudman, L.A. and Glick, P. (2001) 'Prescriptive Gender Stereotypes and Backlash Toward Agentic Women', *Journal of Social Issues*, 57(4), pp. 743–762. Available at: <https://doi.org/10.1111/0022-4537.00239>.
48. Sahoo, N.R. *et al.* (2024) 'Addressing Bias and Hallucination in Large Language Models', in R. Klinger *et al.* (eds) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*. Torino, Italia: ELRA and ICCL, pp. 73–79. Available at: <https://aclanthology.org/2024.lrec-tutorials.12> (Accessed: 14 November 2024).
49. Salikutluk, V. *et al.* (2023) 'Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems', in *HHAI 2023: Augmenting Human Intellect*. IOS Press, pp. 153–167. Available at: <https://ebooks.iospress.nl/doi/10.3233/FAIA230081> (Accessed: 16 November 2024).
50. Salinas, A. and Morstatter, F. (2024) 'The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance'. arXiv. Available at: <http://arxiv.org/abs/2401.03729> (Accessed: 12 August 2024).
51. Schaller, M. (1992) 'In-group favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes', *Journal of Personality and Social Psychology*, 63(1), pp. 61–74. Available at: <https://doi.org/10.1037/0022-3514.63.1.61>.
52. Schmitt, F., Sundermeier, J. and Mikkelsen, K. (2023) 'Hello Diversity! Digital Ideation Hackathon: Fostering Gender Equality in Digital Innovation', *Communications of the Association for Information Systems*, 53(1), pp. 877–902. Available at: <https://doi.org/10.17705/1CAIS.05336>.
53. Schramowski, P. *et al.* (2022) 'Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2103.11790>.
54. Serenko, A. and Turel, O. (2021) 'Why Are Women Underrepresented in the American IT Industry? The Role of Explicit and Implicit Gender Identities', *Journal of the Association for Information Systems*, 22(1), pp. 41–66. Available at: <https://doi.org/10.17705/1jais.00653>.
55. Short, C.E. and Short, J.C. (2023) 'The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation', *Journal of Business Venturing Insights*, 19, p. e00388. Available at: <https://doi.org/10.1016/j.jbvi.2023.e00388>.

56. Smith, E.M. *et al.* (2022) “‘I’m sorry to hear that’: Finding New Biases in Language Models with a Holistic Descriptor Dataset”, in Y. Goldberg, Z. Kozareva, and Y. Zhang (eds) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9180–9211. Available at: <https://doi.org/10.18653/v1/2022.emnlp-main.625>.
57. Strachan, R. *et al.* (2018) ‘Women in engineering: Addressing the gender gap, exploring trust and our unconscious bias’, in *2018 IEEE Global Engineering Education Conference (EDUCON). 2018 IEEE Global Engineering Education Conference (EDUCON)*, pp. 2088–2093. Available at: <https://doi.org/10.1109/EDUCON.2018.8363497>.
58. Sun, Z. (2023) ‘A Short Survey of Viewing Large Language Models in Legal Aspect’. Available at: <https://doi.org/10.48550/arXiv.2303.09136>.
59. *The state of AI in early 2024 | McKinsey* (no date). Available at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai/> (Accessed: 16 November 2024).
60. Turner, J.C., Brown, R.J. and Tajfel, H. (1979) ‘Social comparison and group interest in ingroup favouritism’, *European Journal of Social Psychology*, 9(2), pp. 187–204. Available at: <https://doi.org/10.1002/ejsp.2420090207>.
61. Ughetto, E. *et al.* (2020) ‘Female entrepreneurship in the digital era’, *Small Business Economics*, 55(2), pp. 305–312. Available at: <https://doi.org/10.1007/s11187-019-00298-8>.
62. Upadhyay, N. *et al.* (2022) ‘The influence of digital entrepreneurship and entrepreneurial orientation on intention of family businesses to adopt artificial intelligence: examining the mediating role of business innovativeness’, *International Journal of Entrepreneurial Behavior & Research*, 29(1), pp. 80–115. Available at: <https://doi.org/10.1108/IJEBR-02-2022-0154>.
63. Wan, Y. *et al.* (2023) “‘Kelly is a Warm Person, Joseph is a Role Model’: Gender Biases in LLM-Generated Reference Letters”. arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.09219>.
64. Wright, M., Siegel, D.S. and Mustar, P. (2017) ‘An emerging ecosystem for student start-ups’, *The Journal of Technology Transfer*, 42(4), pp. 909–922. Available at: <https://doi.org/10.1007/s10961-017-9558-z>.
65. Wu, S. *et al.* (2023) ‘BloombergGPT: A Large Language Model for Finance’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2303.17564>.
66. Wu, Y. *et al.* (2023) ‘Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.09241>.
67. Zhao, J. *et al.* (2018) ‘Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods’, in M. Walker, H. Ji, and A. Stent (eds) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). NAACL-HLT 2018*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20. Available at: <https://doi.org/10.18653/v1/N18-2003>.