Marsden, Nicola, Rothe-Wulf, Annelie, Tang, Yvonne, Herling, Claudia & Reichert, Tim. (2025). Navigating Bias: Using LLMs to Analyze Discrimination in Entrepreneurial Game Dialogues. HCI International, LNCS 15816, 202–219.
doi:10.1007/978-3-031-92578-8_14

Navigating Bias: Using LLMs to Analyze Discrimination in Entrepreneurial Game Dialogues

Nicola Marsden, Annelie Rothe-Wulf, Yvonne Tang, Claudia Herling, and Tim Reichert

Heilbronn University, Heilbronn, Germany
{nicola.marsden,annelie.rothe-wulf,yvonne.tang,claudia.herling,tim.reichert}@hs-heilbronn.de
https://www.hs-heilbronn.de/de/lab-sozioinformatik

Abstract. Entrepreneurship is a critical driver of innovation, yet women remain significantly underrepresented in the field due to persistent gender biases that restrict access to resources, funding, and networks. This paper presents insights from developing a serious game that combines decision-tree narratives with Large Language Models (LLMs) to simulate entrepreneurial scenarios and provide real-time bias detection and feedback. Players navigate interactive dialogues that reflect common gender biases in entrepreneurship, fostering awareness and equipping them with strategies to handle discriminatory situations. The game integrates a domain-specific knowledge base of over 20 biases with structured LLM analysis to ensure accurate detection and tailored feedback. Developed through an iterative co-design process involving women entrepreneurs, consultants, and gender experts, the game emphasizes emotional safety and contextual relevance. Our findings from user testing and expert evaluations demonstrate how this approach fosters resilience and preparedness. This research contributes to gender equity in entrepreneurship and AI-assisted educational tools, advancing inclusive and bias-aware innovation ecosystems.

Keywords: Gender bias · Serious games · Entrepreneurship · AI and ethics.

1 Introduction

Entrepreneurship is a significant driver of innovation and economic growth, yet gender biases continue to hinder women's participation by restricting access to resources, funding, and networks [22,23]. These biases manifest throughout the entrepreneurial journey, including during business formation, negotiations, and decision-making, often through undervaluation, leadership stereotypes, and assumptions about risk-taking [13]. Addressing these barriers requires tools that promote awareness and foster resilience, and serious games—particularly those leveraging visual novel mechanics—have proven effective at engaging users in reflective, scenario-based learning [5,17].

Advances in artificial intelligence (AI), particularly Large Language Models (LLMs), offer new opportunities to address gender bias in entrepreneurship. LLMs can analyze interactions to uncover implicit biases and provide real-time feedback, enhancing users' awareness of discriminatory dynamics in professional contexts. However, LLMs are prone to replicating biases embedded in their training data [10,24], underscoring the need for mitigation strategies such as prompt engineering and self-debiasing techniques [28,29].

This paper builds on our work on a serious game that combines decision-tree-based narratives with real-time bias detection powered by an LLM to empower women entrepreneurs. The game immerses players in interactive dialogues reflecting gender bias scenarios, supported by a domain-specific knowledge base of over 20 gender biases. Through structured LLM feedback, players develop strategies to navigate these biases effectively. The system ensures emotional safety by constraining the LLM's role to predefined scenarios, reducing risks associated with generative AI outputs [3].

The game was developed using an iterative co-design process involving women entrepreneurs, consultants, and gender experts, ensuring relevance, inclusivity, and emotional safety. In [19], a technical overview of the game's architecture, a prototypical approach to LLM-based novel creation, and an initial evaluation of LLM-based feedback are presented. This paper concentrates on the detection of bias and discrimination and provides a deeper exploration of the co-design process employed in the project. To situate our approach, we first explore existing research on social bias in AI and the role of serious games in fostering awareness and resilience.

2 Related Work

2.1 Social Bias in LLMs

Large Language Models have revolutionized natural language processing by generating human-like text and assisting with various tasks, but they are not without flaws. A significant issue is that LLMs often reproduce societal and historical biases present in their training data [10]. These biases can manifest in various ways, from stereotypical language to discriminatory assumptions embedded in generated outputs [3]. Recognizing these risks, researchers have proposed a variety of bias mitigation techniques to make LLMs fairer and more equitable in their responses. Various debiasing techniques—spanning dataset augmentation, fairness constraints during training, and output adjustments—aim to reduce biases in LLMs [29,28,7].

Self-debiasing techniques have recently gained attention as they enable models to identify and adjust their own biases without additional external intervention. Isabel Gallegos et al. [10] highlight zero-shot self-debiasing methods, where an LLM is prompted to recognize potential stereotypes and adjust its outputs accordingly. This method is efficient, scalable, and adaptable across various applications, making it particularly valuable for interactive systems like serious games.

Despite the advancements in debiasing techniques, challenges remain in ensuring cultural adaptability and intersectional fairness. Most bias mitigation methods have been developed with a focus on Western contexts, which may not fully capture the nuances of biases in other cultural settings [15]. Addressing intersectional biases—where gender intersects with race, age, disability, and other identities—requires more sophisticated approaches to bias detection and correction [2].

2.2 Serious Games for Social Awareness

Serious games have been recognized as a powerful medium for promoting social awareness and behavioral change. These games create immersive, interactive environments where players can explore complex issues in a safe, controlled manner. Research shows that serious games can effectively foster empathy, enhance understanding of social issues, and encourage reflection on personal biases [17,11]. Utilizing positive framing and enjoyable experiences, serious games can engage players in meaningful social issues, encouraging them to reflect on their attitudes and behaviors [26].

Visual novels are a type of game in which players are presented with narrative-driven scenarios where they must make choices that influence the story's outcome. Studies have shown that real-life scenarios and the ability to share personal experiences enhance players' reflection and learning [5]. Visual novels provide a unique opportunity to simulate real-world challenges and encourage players to experiment with different responses to discriminatory situations [12]. In visual novels, characters interact with the player directly and involve them in a dialogue in which the player can choose different options [6]. Players can interact with characters in the first person to improve player identification and immersion [4]. This first-person perspective means that players do not need to be presented by an avatar, which allows for faster onboarding since the player does not have to select or customize an avatar. It also means that the problematic implications of stereotype threat through the avatar is reduced for marginalized identities. Stereotype threat is a phenomenon which leads to people acting in line with the stereotypes attributed to their group when their group identity is made salient, e.g., Black people having a worse performance on math tests when a question of the participants' race preceded the assessment [21]. In games involving avatars, this can occur because the player's avatar's characteristics may influence the player's communication [18,27]. Accordingly, for players whose social identities are subject to stereotypization, avatar embodiment can enact stereotype-consistent behaviors, and this can have a negative effect on marginalized identities.

These findings informed the design of our serious game, which leverages AI for bias detection while prioritizing narrative control and emotional safety.

3 Designing a Bias-Responsive Serious Game

Our game addresses two critical challenges: the gender biases faced by women entrepreneurs and the biases inherent in LLMs. By integrating manually crafted decision-tree dialogues with real-time LLM feedback, we provide players a controlled environment to recognize and reflect on biases. This combination ensures realistic, culturally relevant scenarios while minimizing risks associated with AI-generated outputs. The following sections detail our game's concept, architecture, dialogue structure, and bias detection mechanisms.

3.1 Game Concept and Objectives

The game uses visual novels to immerse women entrepreneurs in realistic, interactive scenarios that highlight gender biases. Players navigate branching dialogues, choosing responses to biased remarks in funding talks, media interviews, and family interactions. This narrative-driven format fosters reflection on real-world biases from multiple perspectives [5]. The game's core objective is to help players

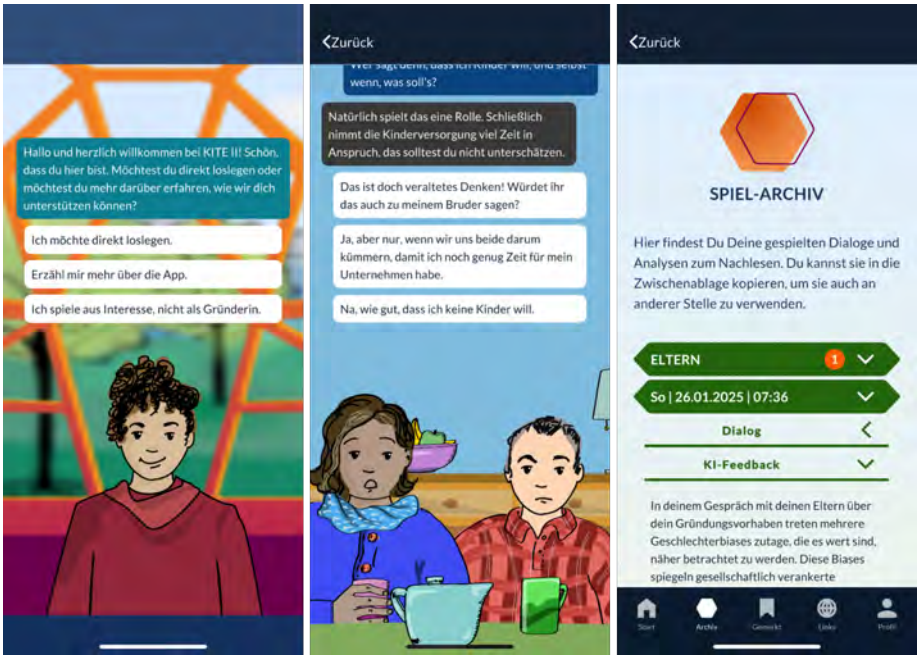

Fig. 1. Prototype screenshots of welcome dialog (left), visual novel example (center), and game archive (right)

recognize and address gender biases in entrepreneurship. Manually crafted dialogues simulate common challenges identified in research [22], ensuring relevance and cultural accuracy.

A distinctive feature is the integration of LLMs for real-time bias analysis. The LLM detects instances of bias during gameplay and provides players with feedback to encourage reflection and exploration of alternative responses [9,24]. The game uses manually crafted decision-tree dialogues to ensure emotional safety, maintaining narrative control and reducing risks associated with generative AI [3]. This approach provides a structured and safe environment for players to explore scenarios, emphasizing learning over judgment. Combining storytelling with AI feedback creates a practical tool to build awareness and resilience in women entrepreneurs.

Fig. 1 displays three screenshots from the current game prototype. The left screenshot features a welcome dialog with a recurring character who acts as a guide throughout the game, offering players the choice to proceed or request additional information. The center screenshot showcases a dialog from a novel where players interact with parents questioning the decision to start a company. The right screenshot highlights the game archive, where the protocol of all game sessions is stored, along with the generated LLM-based feedback.

3.2 Game Architecture

The serious game is designed with a modular architecture to facilitate seamless integration with various LLMs. This design ensures adaptability and scalability, allowing the system to incorporate advancements in AI technology without significant structural changes.

Dialogue System

- **Structure:** The game employs a branching dialogue system, utilizing manually crafted dialogue trees to guide player interactions. This approach maintains narrative coherence and emotional safety, preventing unintended content generation.
- **Functionality:** Players navigate through predefined dialogue options, making choices that influence the storyline and character development. This system ensures that interactions are meaningful and aligned with the game's educational objectives.

Bias Detection Module

- **Integration with LLMs:** The game incorporates LLMs to analyze player interactions within the dialogue system. By processing dialogue logs, the LLM identifies instances of gender bias or discrimination, providing real-time analysis based on a predefined knowledge base of biases.
- **Prompt Engineering:** To ensure accurate bias detection, the system utilizes structured prompts divided into components such as Role, Assignment, Context, Knowledge Base, and Output Format. This structured approach guides the LLM's analysis, reducing the risk of irrelevant outputs.

Feedback Mechanism

- **Real-Time Feedback:** Upon detecting biased interactions, the system delivers immediate, context-specific feedback to the player. This feedback encourages reflection and offers alternative responses, enhancing the learning experience.
- **Emotional Safety:** The feedback is designed to be non-judgmental and supportive, fostering a discrimination-sensitive environment for players to explore sensitive topics without fear of criticism. This approach promotes engagement and openness to learning.

This modular architecture not only supports the current integration of LLMs but also allows for future enhancements, such as incorporating more advanced AI models or expanding the game's content to address a broader range of biases and scenarios. The design ensures that each component functions cohesively, providing a robust platform for educating players about gender biases in entrepreneurship.

3.3 Game Design and Dialogue Structure

The game's narrative structure is built using branching dialogue trees within preconstructed visual novels. These dialogues were developed in collaboration with experts in gender studies and entrepreneurship to reflect real-world scenarios women entrepreneurs frequently encounter [20]. The game currently presents six key scenarios, with each scenario covering a variety of biases:

- **Renting office space:** The property owner suggests that older founders are merely pursuing a hobby.
- **Appointment with a notary:** The notary assumes that the company's head must be a man.
- **Negotiating contract fees:** The negotiating partner insinuates that women are more emotional during negotiations.
- **Informing parents about becoming an entrepreneur:** The parents express doubts about the player's entrepreneurial capabilities.
- **Being interviewed by a journalist:** The press focuses on the player's gender rather than the business idea.
- **Appointment for a loan:** The bank employee inquires about family planning, implying potential interference with business commitments.

Currently, these dialogues are manually crafted using Twine, an interactive storytelling tool, and imported into Unity for integration within the game. Players choose from various dialogue options that affect the course of the story, resulting in a log of concrete interactions. The LLM analyzes this log to provide personalized feedback, helping players recognize the biases they encountered and get feedback on their reactions. Central to ensuring the accuracy and reliability of this feedback is the development of structured prompts, which guide the AI's analysis and responses.

4 Prompt Engineering

Prompt engineering in the context of our app enables accurate bias detection and tailored feedback by combining a domain-specific knowledge base with structured prompts. This approach ensures the LLM identifies biases, analyzes player behavior, and delivers constructive, emotionally safe feedback, supporting players in navigating gender biases in entrepreneurial scenarios.

4.1 Knowledge Base of Biases

The bias detection mechanism in the game is supported by a domain-specific knowledge base that includes over twenty biases relevant to women entrepreneurship [13,16,22,25]. These biases encompass financial biases, maternal biases, and stereotypes related to leadership and risk-taking. The knowledge base provides detailed descriptions of each bias, ensuring that the LLM can accurately detect and provide context-specific feedback during gameplay.

For instance, in one scenario, the player may encounter a bank employee asking about their family plans during a loan negotiation, reflecting a maternal bias that assumes women with children are less committed to their businesses. Another scenario might involve a property owner questioning the player's financial risk-taking abilities, illustrating a stereotype about women's risk aversion. By incorporating these biases into the game's knowledge base, the system ensures that feedback is precise and relevant to real-world experiences.

The game also addresses intersectional biases, such as heteronormativity and age-related stereotypes, to point to the challenges faced by women from marginalized backgrounds. The knowledge base is regularly updated through co-design workshops with gender experts and women entrepreneurs to remain contextually accurate and culturally relevant. This iterative process ensures that the game remains a valuable tool for educating players on navigating gender biases in various entrepreneurial settings.

4.2 Prompt Structure and Bias Detection

Effective bias detection within the KITE II game environment relies heavily on structured prompt engineering and a modular system design. The prompts used to guide the LLM analysis are crafted with precision to ensure the system accurately identifies biases in dialogues while minimizing hallucinations and false positives. The bias detection mechanism in KITE II combines the structured prompt engineering approach with self-debiasing techniques to enhance the accuracy and relevance of the feedback provided to players. The system employs self-debiasing via explanation as the primary debiasing method [10]: The LLM is prompted to identify potential stereotypes or biases before generating the analysis. For this, the structured prompt engineering approach used in the game follows a five-part format [19]: Role, Assignment, Context, Knowledge Base, Output Format, and Target. The LLM analyzes the dialogue logs in real time,

Table 1. Structure of the prompt for discrimination detection and feedback, adapted from [19]

Section	**Example**
Role	"You are a gender researcher."
Assignment	"Your assignment is to analyze the log for discrimination and gender biases and give feedback to the player..."
Context	"The founder plans to start a software company and talks to her concerned mother about it."
Knowledge Base	Consisting of a description of biases that women founders are confronted with.
Output Format	Target "Write an analysis of the dialog..."
Target	This entails the log of dialogue (Mother: "Hello and thank you for coming" Player: "Hi"...)

comparing player responses and dialogue paths to the domain-specific knowledge base of biases.

This modular approach allows for flexibility and scalability, enabling updates to the knowledge base as new biases or cultural contexts are identified. It also ensures that the LLM can be adapted to different entrepreneurial scenarios and dialogue structures without requiring significant reengineering.

4.3 Feedback Structure and Example

The feedback for the player starts with a short summary and general assessment of the player's performance. The following is an example of such a section generated for a playthrough of the loan appointment novel: "In your conversation with the bank employee, several gender-related biases emerged that needed to be recognized and navigated. You prepared well and handled the challenges confidently. Let's take a closer look at the biases and your reactions to them."

Then a recognized bias is listed and explained, for example in the case of the loan appointment novel: "Undervaluation of Women-Led Businesses and Gender-Specific Stereotypes: The bank employee expressed doubts (...). These doubts reflect the undervaluation of women-led businesses and gender-specific stereotypes, which often portray women as less competent or less willing to take risks."

The system then analyzes the player's reaction, giving an assessment of potential advantages and disadvantages. Again, for the bank loan example, the player's reaction is described as a "confident defense" based on a "realistic assessment" of the founding situation. Advantages like "composure and professionalism" and a "focus on facts" are given. As a potential drawback, the system notes that through "avoidance of direct confrontation," biases "remain unchallenged."

The feedback ends with a short conclusion, stating that the player "demonstrated strength and clarity in a potentially challenging situation."

4.4 Additional Instructions for Emotional Safety

The prompt ensures emotional safety through specific instructions to the LLM:

1. Use provided clues for biases to analyze the dialog.
2. Analyze the player behavior and reactions to the biases with concrete examples from the dialog.
3. Highlight advantages of the player behavior and cautiously suggest what drawbacks the reactions might have.
4. Do not list the non-addressing of gender-stereotypical assumptions as a drawback.
5. Be cautious about directly addressing biases and stereotypes, as this can be generally helpful, but the player's primary focus should be on conducting the conversation successfully.
6. Use gender-inclusive language and be supportive and encouraging.

The first instruction ensures that the system does not overlook obvious biases, while the second instruction aims for feedback that is concrete and comprehensible in the context of the game session. The fourth and fifth instruction emphasize that the goal is not for the player to directly address inappropriate behavior but to successfully navigate the founding-related situation.

The continuous refinement of these prompts was integral to our co-design process, ensuring that the system's outputs aligned with user needs and expectations.

5 Iterative Co-Design and Evaluation

The iterative co-design process employed in developing the KITE II serious game is grounded in the user-centered design principles outlined by ISO 9241-210. This methodology structures the design and development process into four phases: understanding and defining the usage context, specifying user requirements, creating design solutions, and evaluating the effectiveness of these solutions [1,14]. The approach prioritizes active stakeholder participation to ensure that the developed solution aligns with real-world needs and challenges. The process is iterative, involving continuous feedback loops that integrate user feedback and expert evaluations. It emphasizes collaborative workshops to refine game content, ensure contextual relevance, and mitigate biases in dialogue interactions.

The KITE II project applied a series of structured workshops and evaluations synchronized into the iterations to ensure the serious game's effectiveness in raising awareness of gender biases. These sessions were critical to developing the game's narrative framework, refining the integration of LLMs, and validating the bias detection mechanism.

5.1 Co-Design Workshops

The co-design workshops were conducted in multiple phases, engaging stakeholders such as coaches of women entrepreneurs, consultants, gender researchers, and representatives from the National Agency for Women Startup Activities and Services in Germany—the number of participants varied between 10 and 20.

These workshops ensured that the game content was grounded in the real-world challenges faced by women entrepreneurs, maintaining contextual relevance and fostering emotional safety.

The first workshop in December 2023 laid the foundation for the game's narrative structure and integration of LLMs. Participants explored various formats, eventually selecting visual novels with decision-tree dialogues due to their ability to offer a structured, emotionally safe environment for players to engage with gender bias scenarios. This workshop also established the decision to integrate LLMs for real-time bias detection, with an emphasis on balancing AI analysis with human-curated narratives to maintain control over sensitive content.

The subsequent workshops in February and March 2024 focused on prompt engineering to guide the LLM's analysis of dialogues. Participants in these workshops developed a structured prompt framework consisting of six key components: Role, Assignment, Context, Knowledge Base, Output Format, and Target. These components ensured that the AI analysis was tailored to the specific context of each dialogue and minimized the risk of hallucinations. Detailed discussions were held on how to structure the prompts to accurately reflect the content of manually crafted dialogues, ensuring that the LLM-generated feedback remained relevant and contextually appropriate.

A workshop in August 2024 prioritized inclusivity and intersectionality. Collaborating with diversity and inclusion experts, stakeholders expanded the narrative scenarios to reflect diverse family structures, cultural contexts, and intersectional challenges, including those faced by non-binary and older entrepreneurs. This workshop also reviewed emotional safety measures to ensure that players could explore complex topics without fear of judgment or harm.

The co-design workshops were pivotal in shaping both the narrative framework and the technical implementation of the game, ensuring that all design decisions were informed by expert input and stakeholder experiences. The key outcomes include:

- **Narrative Alignment with Real-World Challenges:** The decision-tree dialogues were crafted to mirror common biases encountered by women entrepreneurs, ensuring contextual accuracy and emotional safety in gameplay.
- **Prompt Engineering Strategy:** Stakeholders contributed to refining the prompt framework to improve the accuracy of the LLM's bias detection. This process focused on structuring prompts to align with real-world entrepreneurial challenges faced by women, reducing the likelihood of AI hallucinations.
- **Inclusivity and Representation:** Discussions around the game's scenarios highlighted the need to incorporate diverse family constellations and intersectional identities. As a result, the narrative paths were expanded to reflect a broader spectrum of experiences.

5.2 Expert Evaluation

An expert evaluation was conducted in January 2024 to validate the accuracy and reliability of the game's bias detection mechanism (cf. [19]. Four domain ex-

perts in women entrepreneurship and gender studies reviewed the decision-tree dialogues and corresponding AI analyzes to ensure that the feedback provided to players was both accurate and context-specific. This evaluation phase led to improvements in the narrative elements and further refinement of the LLM prompts, ensuring that the AI analysis aligned with the intended scenarios without introducing extraneous or irrelevant outputs.

The experts assessed the system's ability to detect predefined biases relevant to women entrepreneurship, such as maternal bias and performance attribution bias. Their feedback played a critical role in refining the prompt structures to ensure consistent and context-relevant analysis by the LLM. Additionally, they provided guidance on managing potential conflicts between AI interpretations and human-crafted dialogues, emphasizing the importance of limiting the LLM's scope to scenarios explicitly marked as containing biases.

The expert evaluation provided critical feedback on the accuracy and relevance of the bias detection mechanism. The key insights from this phase include:

- **Accuracy of Bias Detection:** Experts found that the LLM's ability to identify specific biases in dialogue paths was generally accurate, though improvements were needed to fine-tune the prompts further. Several biases were correctly flagged by the AI, while others required adjustments to the knowledge base and prompt structure.
- **Contextual Relevance:** The experts noted that the game effectively captured common discriminatory scenarios faced by women entrepreneurs, but they recommended further contextual refinement to ensure that the AI analysis remained relevant across various cultural and situational contexts.
- **Prompt Refinements:** Based on expert feedback, several prompts were revised to clarify the expected responses from the AI, enhancing the system's interpretive accuracy.

5.3 User Testing

User testing sessions were conducted in July 2024, focusing on the usability and navigability of the game's dialogue system and AI analysis. Eleven participants, ranging in age and professional backgrounds, engaged with the game scenarios between one and three hours, providing feedback on their experiences. These sessions followed a think-aloud protocol, where participants verbalized their thoughts as they navigated the visual novels. The feedback from these sessions informed iterative updates to the game's design, ensuring that the dialogue options were expanded and the bias detection system became more user-friendly. Key insights included insights into the usability of the game as well as the players' interactions with both the narrative content and the AI analysis:

- **User Engagement and Navigation:** Participants reported that the visual novels were engaging and easy to navigate. However, some users expressed confusion regarding the purpose of certain dialogue options, prompting adjustments to make the objectives clearer.

- **Feedback on Bias Analysis:** Players appreciated the real-time feedback provided by the LLM. However, they suggested that the AI's feedback could be made more personalized and actionable. Users wanted the feedback to focus more specifically on their reaction to the biases they were confronted with.
- **Dialogue Options:** The testing revealed that users desired a wider range of response options within the dialogues. The feedback led to the inclusion of additional decision paths to provide more nuanced interactions.

5.4 Validation of Bias Detection Mechanism

The validation of the bias detection mechanism was an integral part of the development process. The project team employed a structured approach to ensure that the LLM's analysis accurately aligned with the manually crafted dialogue content. The validation process involved repeated testing and refinement of prompts (see section 4), ensuring that the LLM's outputs were relevant and context-specific.

Metrics were established to evaluate the accuracy and consistency of the bias detection system across various scenarios. This process involved ensuring that the LLM detected only the biases explicitly marked within the dialogues, testing the system across a wide range of dialogue paths to ensure that the feedback remained reliable and consistent, and minimizing hallucinations by restricting the LLM's scope to predefined dialogue elements.

The validation process aimed to assess the accuracy, consistency, and reliability of the LLM's bias detection mechanism across various game scenarios.

- **Accuracy Metrics:** The validation phase involved comparing the AI-generated outputs with expert evaluations to measure accuracy. The results showed a high alignment between the LLM's analysis and the experts' assessments, confirming the reliability of the bias detection system.
- **Consistency Across Scenarios:** The system demonstrated consistency in identifying biases across different narrative paths. This finding confirmed that the structured prompt framework effectively guided the LLM's analysis, reducing the occurrence of hallucinations and irrelevant outputs.
- **Iterative Refinements:** Based on validation feedback, the prompts were iteratively refined to further improve the bias detection mechanism. These refinements included adjusting the language used in prompts and incorporating additional context-specific information to enhance the AI's interpretive capabilities.

The iterative development process and stakeholder involvement provided critical insights, which are reflected in the findings discussed in the following section.

6 Findings and Discussion

The iterative development process of KITE II revealed insights into the design, implementation, and evaluation of the game. This section highlights the core

findings from co-design workshops, expert evaluations, and user testing, with a focus on their impact on narrative development, technical innovation, and user interaction. We also discuss the effectiveness of LLM integration and address limitations to guide future improvements.

6.1 Insights from Co-Design and Evaluations

The co-design and evaluation processes conducted during the development visual novel provided critical insights that shaped both the narrative framework and the technical integration of the bias detection mechanism. This section synthesizes the key findings from co-design workshops, expert evaluations, user testing, and the validation process to highlight the impact of these efforts on the game's overall design and functionality.

Enhancing Narrative Control and Emotional Safety Co-design workshops highlighted that manual narrative control enhances emotional safety, ensuring scenarios are pre-approved and aligned with ethical standards. This approach prevents unintended outputs from the LLM, creating a safe environment for exploring discriminatory scenarios.

Improving Prompt Engineering for Bias Detection The iterative prompt engineering process revealed the necessity of a structured framework to guide the LLM's analysis. The workshops and expert evaluations demonstrated that carefully crafted prompts could significantly reduce the risk of hallucinations and irrelevant outputs. The inclusion of contextual elements in prompts, such as specifying the role of the speaker or the context of the dialogue, improved the AI's ability to accurately identify biases within the game's narratives. Based on user testing, which revealed that players desired feedback evaluating their chosen reactions, we revised the prompt to incorporate an analysis of both the benefits and the potential drawbacks of the selected response.

Addressing Usability and Interaction Challenges User testing sessions highlighted several usability challenges that influenced subsequent design iterations. Players found the game engaging but expressed the need for clearer instructions regarding the purpose of the dialogue options and the AI feedback. In response, the development team refined the user interface and adjusted the narrative flow to make the game's objectives more transparent. These changes enhanced player engagement and ensured that users could fully benefit from the game's educational aspects.

Validating the Bias Detection Mechanism The validation process confirmed the reliability and consistency of the bias detection mechanism. The system showed high accuracy in identifying biases across various narrative paths,

aligning closely with expert evaluations. This consistency demonstrated the effectiveness of the structured prompt framework and highlighted the importance of iterative testing to improve the AI's performance. The findings underscored the need for ongoing refinements to maintain the system's accuracy as new scenarios and biases are introduced.

Broadening Inclusivity and Representation The co-design workshops and user testing sessions also provided valuable insights into the need for more inclusive narrative scenarios. Participants recommended expanding the game's dialogues to reflect diverse family constellations, cultural contexts, and intersectional identities. These recommendations led to the inclusion of broader narrative paths that better capture the varied experiences of women entrepreneurs. The findings emphasized that achieving inclusivity in serious games requires continuous engagement with diverse stakeholders.

6.2 Summary of Insights

Overall, the evaluations revealed that the collaborative development process effectively balanced narrative control with technical innovation. The insights gained from stakeholders and users helped refine the game's design, ensuring that it provided a meaningful and context-relevant learning experience for women entrepreneurs. The project demonstrated that integrating LLMs into serious games can be both impactful and ethical when guided by structured prompts, expert input, and iterative validation.

6.3 Effectiveness of LLM Integration

A cornerstone of the project was maintaining manual control over dialogue narratives while employing LLMs solely for bias detection. This approach balanced emotional safety—crucial when simulating potentially distressing or triggering content—with the analytical power of AI. By restricting the LLM's role to assessing player-selected dialogue paths, the team effectively minimized the risk of "hallucinated" or harmful AI-generated responses [3].

Structured prompt engineering emerged as another crucial factor in harnessing the LLM's capabilities. Dividing prompts into Role, Assignment, Context, Knowledge Base, and Output Format reduced misclassifications while also helping the model focus on a clearly defined set of 25 biases. Expert evaluations conducted in early 2024 confirmed that carefully framing prompts yielded consistent, reliable outputs. In particular, domain-specific tagging of relevant dialogue lines deterred the model from over-identifying biases in neutral statements (e.g., offering a coffee). However, the reliance on structured prompts raises questions about the scalability of the approach across different cultural contexts. Future iterations of the system should explore adaptive prompt mechanisms to account for diverse socio-cultural nuances.

6.4 Limitations

While the findings demonstrate the potential of our approach, we also identified areas for improvement, which we discuss in the next section. The current implementation of the KITE II game focuses on gender biases commonly experienced by white female founders in Germany, which may limit its applicability in other cultural contexts. The knowledge base and scenarios reflect challenges specific to this demographic, potentially overlooking biases and systemic barriers that affect entrepreneurs from diverse racial, cultural, or socio-economic backgrounds. Additionally, while the game addresses some intersectional biases, such as those related to age and family roles, it does not fully explore other intersecting identities, such as race, disability, or sexual orientation. Expanding the knowledge base to account for these dimensions will require further collaboration with stakeholders representing diverse perspectives.

Another notable limitation is the reliance on manually crafted dialogues for ensuring emotional safety and narrative control. While this approach reduces the risk of unintended content, it constrains scalability, as creating detailed, culturally sensitive storylines demands significant time and expertise. Moreover, the bias detection mechanism depends heavily on structured prompts and predefined biases, which, while effective, could lead to misinterpretation in more nuanced or unanticipated contexts. Ongoing refinement of prompts and validation processes is essential to maintain the system's reliability. Future iterations should explore integrating adaptive learning mechanisms and automated story generation while retaining the necessary safeguards to ensure both inclusivity and emotional safety.

Addressing these limitations opens avenues for future research, particularly in enhancing scalability and inclusivity.

7 Future Work

To enhance the inclusivity and effectiveness of the KITE II game, future efforts will focus on expanding the knowledge base to address intersectional biases that go beyond gender, such as those related to race, age, disability, and cultural context. Additionally, the game could benefit from integrating adaptive learning loops that dynamically refine the bias detection system based on player interactions and feedback, improving its accuracy and adaptability over time [10].

Further development will also explore methods for scaling the game's content without compromising emotional safety. This includes investigating automated dialogue generation techniques supported by rigorous filtering and validation processes to maintain narrative quality [8,10]. Long-term impact assessments are also necessary to evaluate the game's influence on players' real-world decision-making and resilience in navigating discriminatory situations. Finally, integrating the game into blended workshop formats, where gameplay serves as a catalyst for in-depth discussions and peer exchange, could amplify its educational impact and foster collective strategies for addressing bias in entrepreneurship.

These efforts will further extend the impact of our work, contributing to more inclusive and equitable innovation ecosystems, as summarized in our conclusion.

8 Conclusion

This paper focuses on specific aspects of our serious game framework, including LLM integration, prompt engineering, and the co-design process we used. It provides detailed insights into these areas, along with their limitations and future directions. Overall, the paper demonstrates the potential of integrating Large Language Models into a serious game framework aimed at empowering women entrepreneurs to recognize and navigate gender bias. Our approach balances manually crafted, decision-tree-based visual novel dialogue–ensuring safe, controlled scenarios–with LLM-driven real-time bias detection. Through extensive co-design workshops, expert evaluations, and user testing, the project has shown that LLMs, when tightly constrained and guided by prompt engineering, can enrich the learning experience without compromising emotional safety. By combining cutting-edge AI with human-centered design, this work lays the groundwork for educational tools that not only confront systemic biases but also inspire collective action for a more equitable future.

Acknowledgments. This work has been partially funded by the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (Bundesministerium für Familie, Senioren, Frauen und Jugend - BMFSFJ) under Grant Number 3923406K04 as part of the project 'KI-Thinktank Female Entrepreneurship II - KITE II' in the Funding Line 'AI for Public Good (KI für Gemeinwohl)'. The responsibility for all content supplied lies with the authors.

The authors would like to thank Rebecca Biebl, Florian Diller, Julian Leidel, and Mergim Miftari for their contributions, which were instrumental in the development of this project.

Ethical Compliance The research was conducted in Germany and line with the ethical guidelines of the German Psychological Society (Berufsverband Deutscher Psychologinnen und Psychologen e. V. (BDP)); data was collected and stored in compliance with EU and German data protection laws (DSGVO).

References

1. Ahmadi, M., Weibert, A., Ogonowski, C., Aal, K., Gäckle, K., Marsden, N., Wulf, V.: Challenges and lessons learned by applying living labs in gender and it contexts. 4th Gender & IT Conference (GenderIT'18) p. 239–249 (2018). https://doi.org/10.1145/3196839.3196878
2. Blanco-Justicia, A., Jebreel, N., Manzanares, B., Sánchez, D., Domingo-Ferrer, J., Collell, G., Tan, K.E.: Digital forgetting in large language models: A survey of unlearning methods. arXiv preprint arXiv:2404.02062 (2024). https://doi.org/10.48550/arXiv.2404.02062

3. Blodgett, S.L., Barocas, S., Daumé Iii, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050 (2020). https://doi.org/10.48550/arXiv.2005.14050
4. Bruno, L.: A Glimpse of the Imaginative Environment, pp. 143–170. CrossAsia-eBooks, Heidelberg (2021). https://doi.org/10.25969/mediarep/17150
5. Camingue, J., Carstensdottir, E., Melcer, E.F.: What is a visual novel? Proceedings of the ACM on Human-Computer Interaction **5**(CHI PLAY), 1–18 (2021)
6. Cotton, T., Shepherd, L.A.: Practising safe sex (t): developing a serious game to tackle technology-facilitated sexual violence. 15th International Conference on Applied Human Factors and Ergonomics pp. 176–186 (2024). https://doi.org/10.54941/ahfe1004777
7. Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., Weston, J.: Queens are powerful too: Mitigating gender bias in dialogue generation. arXiv preprint (2019). https://doi.org/10.48550/arXiv.1911.03842
8. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023). https://doi.org/10.48550/arXiv.2305.14325
9. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. arXiv preprint arXiv:2309.00770 (2023). https://doi.org/10.48550/arXiv.2309.00770
10. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Yu, T., Deilamsalehy, H., Zhang, R., Kim, S., Dernoncourt, F.: Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. arXiv preprint arXiv:2402.01981 (2024)
11. Hammer, J., To, A., Schrier, K., Bowman, S.L., Kaufman, G.: Learning and role-playing games, pp. 283–299. Routledge (2018). https://doi.org/10.4324/9781315637532
12. Iacovides, I., Cutting, J., Beeston, J., Cecchinato, M.E., Mekler, E.D., Cairns, P.: Close but not too close: Distance and relevance in designing games for reflection. Proceedings of the ACM on Human-Computer Interaction **6**(CHI PLAY), 1–24 (2022). https://doi.org/10.1145/3549487
13. Laguía, A., García-Ael, C., Wach, D., Moriano, J.A.: "think entrepreneur-think male": a task and relationship scale to measure gender stereotypes in entrepreneurship. International Entrepreneurship and Management Journal **15**, 749–772 (2019). https://doi.org/10.1007/s11365-018-0553-0
14. Marsden, N., Bernecker, T., Zöllner, R., Sußmann, N., Kapser, S.: Buga:log – a real-world laboratory approach to designing an automated transport system for goods in urban areas. Special Issue: IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC 2018) pp. 72–80 (2018)
15. Mattern, J., Jin, Z., Sachan, M., Mihalcea, R., Schölkopf, B.: Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. arXiv preprint (2022). https://doi.org/10.48550/arXiv.2212.10678
16. Mavin, S., Yusupova, M.: Competition and gender: Time's up on essentialist knowledge production. Management Learning **52**(1), 86–108 (2021). https://doi.org/10.1177/1350507620950176
17. McGonigal, J.: Reality is broken: Why games make us better and how they can change the world. Penguin, New York (2011)
18. Pröbster, M., Soto, M.V., Connolly, C., Marsden, N.: Avatar-based virtual reality and the associated gender stereotypes in a university environment. Euro-

pean Journal of Open, Distance and E-Learning **24**(1), 11 – 24 (2022). https://doi.org/10.2478/eurodl-2022-0002
19. Reichert, T., Miftari, M., Herling, C., Marsden, N.: Empowering female founders with ai and play: Integration of a large language model into a serious game with player-generated content. HCI International **LNCS 14731, Part II**, 69–83 (2024). https://doi.org/10.1007/978-3-031-60695-3_5
20. Schirmacher, A., Bey, K.v.d.: Kite-thinktank female entrepreneurship - fighting discrimination for female entrepreneurs through targeted competence building in recognizing and overcoming patterns of discrimination - internal project report. Project KITE **more info on https://www.kite-bga.de/expertisen/ (in German)** (2022)
21. Steele, C.M., Aronson, J.: Stereotype threat and the intellectual test performance of african americans. Journal of personality and social psychology **69**(5), 797–811 (1995)
22. Tonoyan, V., Strohmeyer, R.: Gender role (in-) congruity and resource-provider gender biases: a conceptual model. International Journal of Gender and Entrepreneurship **13**(3), 225–242 (2021). https://doi.org/10.1108/IJGE-12-2020-0201
23. Veckalne, R., Tambovceva, T.: The importance of gender equality in promoting entrepreneurship and innovation. Marketing and Management of Innovations pp. 158–168 (2023). https://doi.org/10.21272/mmi.2023
24. Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.W., Peng, N.: "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. arXiv preprint arXiv:2310.09219 (2023). https://doi.org/10.48550/arXiv.2310.09219
25. Williams, J.C., Dempsey, R.: What works for women at work: Four patterns working women need to know. NYU Press (2018)
26. Yam, A., Russell-Bennett, R., Foth, M., Mulcahy, R.: How does serious m-game technology encourage low-income households to perform socially responsible behaviors? Psychology & Marketing **34**(4), 394–409 (2017). https://doi.org/10.1002/mar.20
27. Yee, N., Bailenson, J.N.: The proteus effect: Self transformations in virtual reality. Human Communication Research **33**(3), 271–90 (2007)
28. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society pp. 335–340 (2018). https://doi.org/10.1145/3278721.3278779
29. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876 (2018). https://doi.org/10.48550/arXiv.1804.06876