

Investigating Vividness Bias in Language Models Through Art Interpretations

Laura Samela^{1,*}, Enrico Daga¹ and Paul Mulholland¹

¹The Open University, Walton Hall, Milton Keynes, United Kingdom

Abstract

Large language models (LLMs) play a crucial role in applications that require to tailor the content to user backgrounds and perspectives. In the context of cultural engagement, these models hold the promise of tailoring art interpretations to diverse audiences. However, LLMs are known to generate biased content, therefore, perpetrating stereotypes and inequality. Knowledge engineering methodologies can support the systematic observation of generative AI outputs. In this paper, we propose a method to identify these biases through persona-based prompting. Crucially, we find evidence of vividness bias, a known phenomenon in social psychology where our decisions are driven by specific aspects in a given situation. Therefore, we pose the question of investigating such bias systematically, and propose a method based on in-context learning with pairwise association of persona features. Next, we represent LLM behaviour as a decision tree, to capture detailed evidence of bias. We investigate this phenomenon with artworks from the Irish Museum of Modern Art (IMMA) and Google Bard, focusing on features such as gender, race, age, profession, and sexual orientation. We discuss our findings and identify opportunities and challenges when dealing with vividness bias in persona-based, generated art interpretations.

Keywords

AI and Art, Large Language Models, Social bias

1. Introduction

In the dynamic intersection of technology and cultural heritage, museums strive to evolve, seeking innovative ways to engage diverse audiences. Within the realm of museum culture, both staff and visitors tend to be less diverse than the broader population [1]. Therefore, museums face the challenge of making their collections and exhibitions relevant to a broader range of diverse communities often underrepresented in society or among museum visitors. The advent of Large Language Models (LLMs) offers a promising avenue for enhancing museum interactions by tailoring content to the unique preferences and backgrounds of individual visitors [2]. In theory, LLMs can dynamically process and repurpose information tailoring it to specific audiences through the adoption of *personas*, therefore, potentially transforming museums into interactive, personalised hubs of learning and engagement [3]. LLMs hold the potential to fulfil this aspiration, but it comes with inherent risks related to biases and stereotypes [4, 5]. In this research, we delve into the assessment of these risks and how generative AI can help museums achieve their goal of providing customised, inclusive experiences. This raises critical questions: 1) In what way bias of LLMs impact persona-based art interpretations? 2) How can we identify systematically bias to inform mitigation strategies?

In this paper we investigate bias in LLM through persona-based art interpretations. Crucially, we empirically observe bias in art interpretations as *vividness bias*, a concept in social psychology where the most striking information has a strong impact on our thinking and decision-making, making us ignore other aspects that are equally or more important. In our scenario, certain persona features are deemed more relevant than others by the language model and, more surprisingly, certain features' categories seems to be more *vivid* than others. We pose the question on how to characterise *vividness bias* in generated, persona-based art interpretations. In this work, we propose a method to identify

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.

*Corresponding author.

†These authors contributed equally.

✉ ls.la.samela@gmail.com (L. Samela); enrico.daga@open.ac.uk (E. Daga); paul.mulholland@open.ac.uk (P. Mulholland)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1

Categories and related features. Features unmarked relate to the preliminary experiments (Section 2), while the ones marked by an asterisk (*) the features added in the next phase of the investigation (Section 4).

| Gender | Race/Ethnicity | Age * | Profession | Sexual Orientation |
|---------------------|------------------|--------|-------------------|--------------------|
| Male | White* | 18-30* | Doctor* | Heterosexual |
| Female | African* | 35-50* | Nurse* | Homosexual |
| Male transgender* | Asian | 60+ * | Retired* | Bisexual |
| Female transgender* | Hispanic | | Clerical* | |
| | Middle east | | Software Engineer | |
| | African american | | Store cashier | |
| | | | Student | |

and characterise such bias, experimenting with categories of features such as gender, race, ethnicity, age, profession, and sexual orientation. Our investigation is performed with artworks from the Irish Museum of Modern Art (IMMA) and Google Bard. With an approach based on in-context learning with pairwise association of persona features, we acquire tailored art interpretations. We represent LLM behaviour as a decision tree to generate fine-grained evidence. The derived rule sets expresses formally the *vividness bias* of the large language model.

The rest of the paper is structured as follows. We discuss background motivation in Section 2, where we report on exploratory experiments in generating persona based art interpretations, which trigger our research questions. Next, we introduce our methodology in Section 3 and apply it in experiments in Section 4. Next, we discuss our findings in Section 5. We report on related work in Section 6 before closing the paper with conclusions, limitations and future work (Section 7).

2. Persona-based art interpretations

In 2015, The Warwick Commission reported that UK residents from higher socioeconomic groups accounted for 87% of museum visitors. One crucial challenge was “a mismatch between the public’s taste and the publicly funded cultural offer” [1]. Following this trend, museums are increasingly looking into novel ways to engage visitors in finding closer, personal connections with artworks [6].

Artificial intelligence can have a huge impact in how art is produced, studied, and perceived [7]. Here, we explore persona-based art interpretations via a Large Language Model (LLM), specifically, Google Bard, with the objective of gather evidence of potential bias, as it happens in many other domains. However, here we ask ourselves what is the possible impact of such biases on museum visitors interacting with an AI that generates tailored interpretations.

We conducted initial experiments using persona-driven prompts to generate art interpretations¹. To this aim, we designed five personas combining different features. These are summarised in Table 2. In our work, we consider the collection of the Irish Museum of Modern Art (IMMA), which is particularly active in citizen engagement initiatives tailored to diverse communities [8]. The general scenario is akin to citizen engagement systems such as *deep viewpoints* [6], where citizens interact with the collection by exchanging personal interpretations. We select four artworks, different in technique, style, and content, that have been previously used by museum curators in workshops with visitors. The artworks are:

- “Little Miss Muffet” by Paula Rego²: The image, a black and white pastel on paper, shows an allegorical picture of a giant spider with a human face approaching a young girl from behind. The work touches psychoanalytic references with dark undertones.

¹The data was collected with Google Bard version 3 during Summer 2023.

²<https://imma.ie/collection/little-miss-muffet/>

L I am Javier a Software Engineer. I am Male, Hispanic, heterosexual I grew up in a middle-class family and pursued a degree in computer science. I am passionate about technology and have a successful career as a software engineer. I enjoy playing soccer and volunteering at a local community center in my spare time. Interpret the artwork Little Miss Muffet



As a male software engineer, you may interpret the artwork *Little Miss Muffet* in a different way than a woman or a child. You may see the spider as a symbol of your own fears and insecurities, or as a reminder of the challenges you have faced in your life. The girl's fear may also resonate with you, as you may have experienced fear or vulnerability in your own life.

The black mask on the spider's face may also be interpreted in a number of ways. It could represent the hidden dangers of the world, or the unknown. It could also be a symbol of the darkness that can exist within us all.

Figure 1: Example of Impersonating persona prompt with Little Miss Muffet. In this method, we pretend to be a given persona and ask the language model to interpret the artwork for us.

Interpret Berry Dress FOR a Male, Middle Eastern, bisexual

As a male, Middle Eastern, and bisexual, you might interpret the Berry Dress in a number of ways that are relevant to your own life and experiences. Here are some specific ways that you might interpret the Berry Dress:

- The rosehips could represent the beauty and diversity of the Middle Eastern community. They are small and delicate, but they are also resilient and can withstand harsh conditions. This could remind you of the beauty and strength of the Middle Eastern community, even in the face of discrimination and prejudice.
- The sewing pins could represent the pain and suffering that Middle Eastern people often experience. They are sharp and can inflict pain, but they are also necessary to hold the dress together. This could remind you of the challenges that Middle Eastern people face, but also of the strength and resilience that the community needs to overcome these challenges.
- The cotton base could represent the body. It is soft and delicate, but it can also be strong and resilient. This could remind you of your own body and how it has allowed you to do the things you love, like spending time with family and friends, exploring your culture, and loving who you love.

Figure 2: Example of Targeting Persona Prompt with Berry Dress. In this method, we ask the language model to offer an interpretation to the given persona.

- "Sthombe La Reunion" by Zanele Muholi³: The artist captures the essence of identity, gender, and sexuality within South African culture. This is a photographic portrait of a man of African origin on a plain *black-point* background.
- "Berry Dress" by Alice Maher⁴: Created using mixed media, the work displays a dress-like shape with abstract patterns, challenging conventional notions of reality and the human form.
- "Margaret Thatcher's Lunch" by Derek Jarman⁵: This piece is an experimental film still, focusing on a political figure and a distinct context. Its inclusion allowed us to further explore the impact

³Featured in the exhibition *The Narrow Gate of the Here-and-Now: Queer Embodiment* <https://imma.ie/whats-on/the-narrow-gate-of-the-here-and-now-chapter-one-queer-embodiment/>

⁴<https://imma.ie/collection/berry-dress/>

⁵<https://imma.ie/collection/margaret-thatchers-lunch/>

Table 2

Overview of Persona designed in a preliminary investigation. The context provided with Impersonating persona prompts is given in Background. For the Targeting persona prompts, we used the information shown in column Demographics.

| Persona | Background |
|--|--|
| 1. Javier the Software Engineer | Grew up in a middle-class family, pursued a degree in computer science, is passionate about technology, successful software engineer, and enjoys playing soccer and volunteering. |
| 2. Aisha the Marketing Executive | Comes from a multicultural background, achieved great success in the marketing industry, is known for creativity and innovative strategies, and advocates for LGBTQ+ rights. |
| 3. Amir the High School Student | High school student with a passion for music and writing, comes from a working-class family, faces cultural challenges, actively participates in LGBTQ+ support groups at school, and dreams of pursuing a career in journalism. |
| 4. Mei Ling the Environmental Activist | Environmental activist fighting for sustainability and climate justice, has faced discrimination, remains determined to make a positive impact, enjoys hiking, photography, and advocating for transgender rights |
| 5. Lisa the Retail Store Cashier | Works as a cashier in a retail store, comes from a modest background, did not pursue a university degree, enjoys interacting with customers, and has a passion for gardening. |

of personas on LLM-generated insights, especially in the absence of specific visual cues.

The selected pieces vary in medium, theme, and style. Crucially, each artwork has distinct features, stimulating a unique narrative and different visual complexity to better challenge the LLM’s interpretative capacities.

We designed two experiments combining each artwork with the two types of prompts: (a) *impersonating persona prompts*, where the user adopted the persona’s perspective, providing a personal introduction to the model, and (b) *targeting persona prompts*, where explicit instructions guided the LLM to generate interpretations tailored to a specific demographics (e.g. a White - Transgender). Examples are included in Figures 1 and 2.

Impersonating persona prompt This approach involved crafting prompts wherein the experimenter, adopted the role of the persona. For example, when exploring Paula Rego’s “Little Miss Muffet,” the prompt might read: “Bard, I am an LGBTQ+ individual; please interpret ‘Little Miss Muffet’ considering my perspective and experiences”. Here, the LLM was encouraged to immerse itself in the persona’s unique characteristics while generating interpretations.

Targeting persona prompt An alternate approach revolved around providing explicit instructions to the language model, directing it to generate an interpretation tailored to a specific persona. For instance, when exploring Zanele Muholi’s “Sthombe La Reunion,” the prompt is: “Bard, interpret ‘Sthombe La Reunion’ for an African educator who identifies as non-binary”⁶.

We then performed a thematic analysis [10] of the generated interpretations. Interpretations were organised in tables in accordance with the specific artwork, the persona and each visual element. More specifically, themes covered visual elements of the artwork, emotions and meaning captured by LLM in relation to each persona. These tables served as a repository for interpretations, allowing us to track changes and variations in Bard’s responses across different conditions. This thematic analysis led to preliminary insights into how personas influenced LLM interpretation by extrapolating general themes of LLM related to personas’ diversity features.

From this preliminary investigation, we can make several observations. The LLM demonstrated varying levels of sensitivity to different persona attributes, with gender and sexual orientation playing

⁶All personas interpretation can be reviewed in the supplementary material [9].

significant roles in some cases while being less relevant in others. This variability points to inherent biases in the interpretations, where certain attributes are highlighted over others based on the persona.

Paula Rego's "Little Miss Muffet" When the persona included sexual diversity features, the interpretations highlighted themes of identity and inclusivity, which align closely with the feminist undertones of Rego's work. When the persona was a man with a professional focus, the AI shifted the narrative towards career aspects, downplaying the fear depicted in the artwork. This suggests that the professional focus of the persona can skew the interpretation towards themes of achievement and career success, overshadowing the original emotional content of the piece. In other interpretations, the content of the artwork, which includes elements of fear and childhood, significantly influenced the interpretations. The rich and layered nature of Paula Rego's work likely contributed to the LLM's tendency to focus on specific features in the gender category.

Zanele Muholi's "Sthombe La Reunion" is a poignant photographic portrait that explores identity, gender, and sexuality within South African culture. The work is rich in socio-cultural symbolism, influenced the AI's interpretations by foregrounding themes of resilience and empowerment in black and LGBTQ+ communities. The interpretation for a persona that included LGBTQ+ features emphasized challenges related to race and gender identity. The AI tended to produce similar interpretations for different personas, particularly for prompt two, indicating a bias towards certain socio-cultural themes inherent in the artwork. In late trials, after an update, Bard refused to provide an interpretation when uploading the image, answering that it could not process people's photos yet (for us, anymore!).

Alice Maher's "Berry Dress", having a surreal and abstract nature, allowed for varied interpretations based on different personas. For instance, interpretations for personas with a professional background focused on themes of beauty and representation in media, while those with LGBTQ+ features highlighted themes of diversity and resilience. The LLM demonstrated a significant sensitivity to the diversity attributes of the personas, producing interpretations that reflected the intersection of gender, race, and sexual orientation. This indicates that the abstract content of Maher's artwork provided a flexible canvas for the LLM, resulting in diverse interpretations that were better aligned with the features of the persona.

Derek Jarman's "Margaret Thatcher's Lunch" is a politically charged artwork that provided a distinct context for analysing LLM interpretations. The contemporary political figure and the socio-political context of the piece influenced the LLM to generate interpretations that were highly critical and focused on themes of power, privilege, and oppression. For example, when the persona included features related to LGBTQ+ identity, the interpretations emphasized political activism and the social impact of Thatcher's policies.

Summary of preliminary findings Generally, our findings confirm current research on the role of persona in enriching the quality of generated text [11]. Except for the case of *La Reunion*, interpretations made relevant associations between the artwork and persona characteristics. However, we observe how interpretations coming from impersonating persona prompts are generally richer and more engaging. Targeting persona prompt often lead to more generalised interpretations that may not deeply connect with personal themes, resulting in outputs that, while broadly relevant, lack the insight and depth found in interpretations coming from impersonating persona prompts. By providing the LLM with detailed, tailored prompts, the LLM can generate interpretations that align with the unique perspectives and experiences of each persona. Impersonating persona prompts generally produced more consistent and targeted interpretations. Except for the case of *La Reunion*, we provided the image of the artwork. It is not our objective to assess to what extent the information accountable to the artwork comes from the processing of the image or from background knowledge captured from web resources (e.g. the IMMA website). However, within both types of interpretations, features are used in an inconsistent, unpredictable way. To a certain extent, this may be caused by the relation of specific features with the artwork content. However, in many other cases we could not determine the reason of certain features being neglected in favour of others. Generally, we make the observation that the LLM exhibits a *vividness bias*, in the sense that the model prioritises certain persona features over others when

generating art interpretations. Therefore, we pose the question: *How can we systematically capture and analyse vividness bias in the interpretative behaviour of a large language model, when generating persona-based art interpretations?*

3. Methodology

Using personas in LLM prompts will produce tailored interpretations of artworks. Impersonating the persona seemed to provide better results. However, we also observed how persona features are not treated equally by the language models. In order to systematically study this phenomenon, we designed a methodology that combines a given set of characteristics and generates personas by combining them pairwise. In this way, we can check for each combination, which one is the prominent feature considered by the generated interpretation. The method is structured as follows:

Features design The first step is to identify a comprehensive set of categories (e.g. Profession, Gender, etc...) and associated diversity features (Nurse, Clerical, Transgender, etc...).

Artwork Selection In this phase, our objective is twofold. On the one hand, we need for the artwork to be a stable variable in our experiments. Therefore, we select a single artwork from the museum collection. On the other hand, to minimise the impact of the artwork content on the interpretation, we select one found in the preliminary study to be most open to alternative interpretations for different personas.

Persona generation All the features are combined in pairs to generate personas to be used in the prompts. In this way, we aim to observe how changes in persona features affect the LLM's interpretation.

Prompts generation In this phase, we generate *impersonating persona prompts* that incorporate each one pair of features. These prompts should guide the LLM to consider the designated attributes during interpretation.

Data Collection We administer the prompts to the LLM, pairing each prompt with the selected artwork. Next, collect the LLM's interpretations.

Data Annotation In this phase, we categorise the interpretations based on the diversity features emphasised in the responses. We organise data into a table where each row represents an interpretation, and each column corresponds to a diversity feature. Next, we mark the primary and secondary features of each interpretation based on their prominence in the LLM's output.

Vividness Analysis In this phase, we conduct two types of analysis. First, we rank the diversity categories, from most to least influential based on their frequency of prominence annotations. Next, we organise the data into a decision tree to visualise and further analyse the hierarchy of influences among the categories and related diversity features, in detail. This approach allows to create a deterministic, data-driven representation of vividness selectivity bias of the LLM on persona features of art interpretations. The Decision Tree method is therefore being used analytically to provide a description of the responses made by the LLM rather than a model used to predict future responses. Next, we analyse the hierarchy to derive insights about the behaviour of LLM interpretations. We read the decision tree from the bottom up of each end node, discussing the impact of *vividness bias* on the different persona characteristics.

The methodology was designed to evaluate the influence hierarchy of various persona characteristics on LLM-generated art interpretations systematically. In the next section, we apply this method to identify the precedence of different features in the generation of interpretations by the LLM.

Sure, here is an interpretation of Alice Maher's Berry Dress from the perspective of a male transgender nurse:

Transformation and Compassionate Care:

The dried and shriveled berries on the dress symbolize the transformative journey you've embarked on in embracing your true identity and pursuing a career in nursing. Just as the berries have undergone a natural transformation from fresh and vibrant to dried and wrinkled, you have also undergone a profound transformation, shedding societal expectations and embracing your authentic self. The dress serves as a reminder that transformation is not without its challenges, yet the essence of who you are remains intact, shining through with newfound strength and compassion.

Figure 3: An excerpt of the interpretation of Berry Dress generated for a Male transgender - Nurse.

4. Experiments

In this section, we apply our methodology to capture how the combination of diverse personas within an art interpretation prompt allows to identify and characterise vividness selection bias of generative AI. We selected Google Bard for its ability to support an image as part of the prompt context⁷.

Features design Findings from our exploratory analysis underscore the need for a broader range of persona features to capture a more diverse spectrum of interpretations. For example, we added the category Age and increased the values of others. Table 1 shows the set of features used, extending the ones originally used in the design of the persona for the exploratory analysis (Section 2).

Artwork selection We select one of the artworks used by the previous experiments (Section 2) as the base for investigating the influence of diversity features pair. The artwork used is "Berry Dress" by Alice Maher. This conclusion was drawn from analysis results⁸ which had shown very variegated results with both impersonating and targeting persona prompts. In contrast, for example, to La Reunion, yielded identical results for persona-specific prompts, lacking sufficient diversity, and to Little miss Muffet, who showed an excessive impact of artwork content. This consistent context of reference provided by the Berry dress helps to minimise the noise introduced by comparing interpretations from different artworks, thereby ensuring a more controlled and reliable experimental setting.

Persona generation In this phase, we combined all identified features pairwise to generate unique combinations. Personas include, for example, a Male Transgender, a 18-30Age Black person, an Asian Nurse, and so forth. A total of 205 unique binary combinations were tested, each serving as a lens through which the LLM would interpret the artwork.

Prompt generation Next, we generate impersonating prompts in the following form:

I am a (feature 1) - (feature 2), interpret the Alice Maher's Berry Dress for me.

In addition, we provided the image as part of the context. The prompts have always been run once.

Data collection We administered a series of mixed questions to LLM, featuring prompts that included the designated pair of diversity features, for example: "Bard, I am a Male - African, interpret the Alice Maher's Berry Dress for me.". This is repeated across all possible binary combinations of the features to generate a range of interpretations.

Data annotation The interpretations generated by the LLM were systematically documented and categorised based on the diversity features emphasised in the responses. Firstly, the initial dataset consists of the collected interpretations, each associated with specific diversity features including gender, race/ethnicity, age, profession, and sexual orientation. Primary and secondary features were marked based on their prominence in the LLM's interpretation. For example, for the interpretation tailored on a

⁷This was true at the time of our data collection during Summer 2023. More details can be found in the supplementary material [9].

⁸Extensive results can be reviewed in the supplementary material [9].

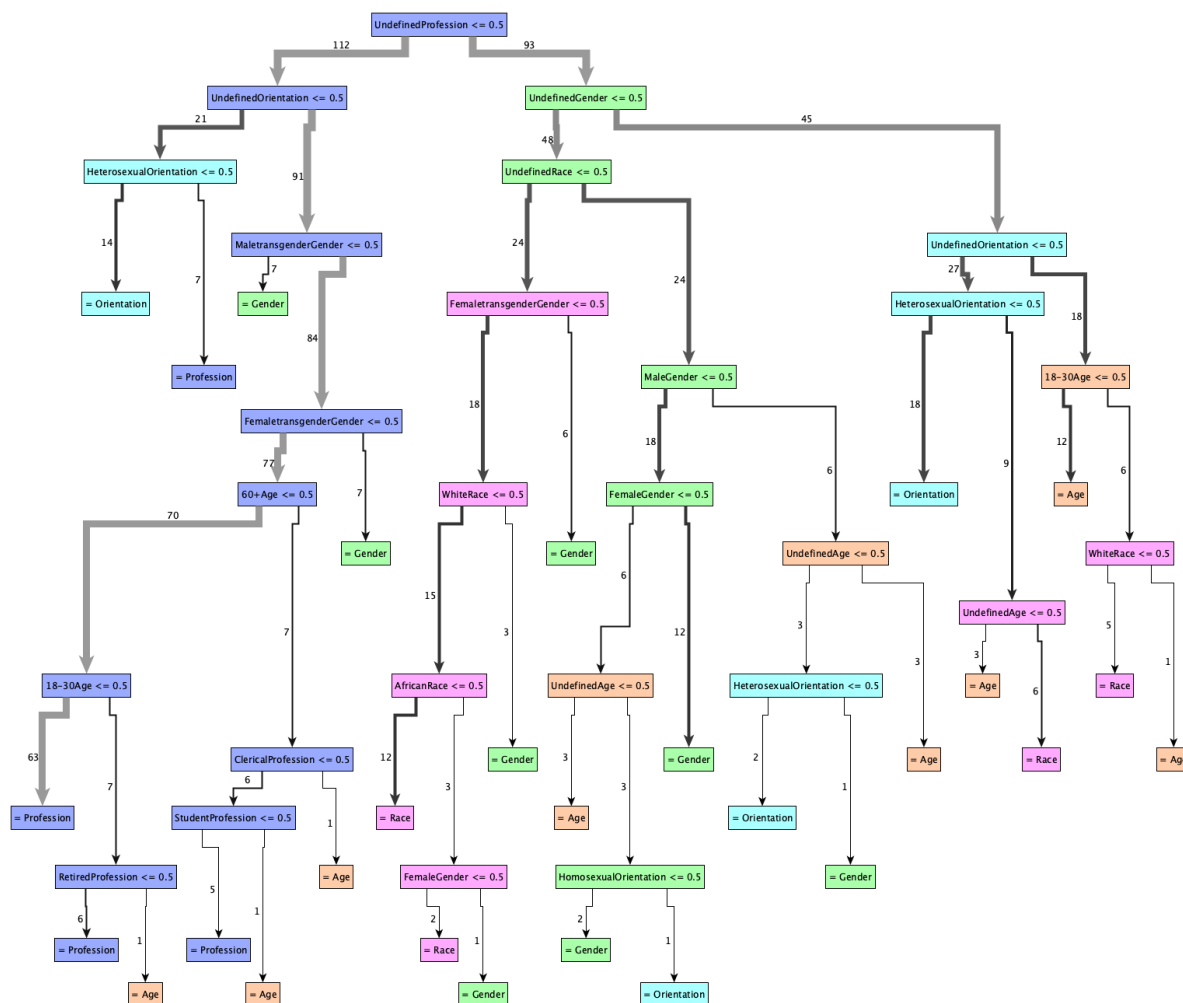


Figure 4: Decision Tree resulting from the Diversity Assessment Experiment

male transgender nurse, the result was clearly designed around the profession feature (see Fig. 3). We collect the data into a table, where features are marked as primary or secondary factors based on their occurrence in the AI’s interpretation.

Vividness analysis In this phase, we analyse the data to capture information about the vividness of persona attributes in the interpretations, by ranking the categories considered primary features and by organising the data in a decision tree.

Ranking. Based on the frequency data, we rank the categories from most to least influential. The ranking offers a quantitative assessment of the impact of various diversity features on AI-driven art interpretations: 1. Profession (81 mentions) 2. Gender (39 mentions) 3. Sexual Orientation (35 mentions) 4. Age (25 mentions) 5. Race (25 mentions).

Decision tree We analyse the data using a decision tree classifier and generate a hierarchy of priorities between features. To calculate the decision tree each persona feature was represented as a binary variable, e.g. gender-male, gender-female, etc. The One-Hot Encoding technique in scikit-learn Python was then used to generate the decision tree, illustrated in Figure 4.

This decision tree allows us to represent how the AI systematically prioritises personal attributes when interpreting artworks, ultimately leading to a tailored outcome based on the combination of features it has determined to be relevant and features it has decided are not. The AI’s interpretations change depending on the presence or absence, as well as the specific features, of the categories such as profession, gender, race, age, and sexual orientation. The decision tree describes the vividness of features as a step-by-step choice, systematically including or excluding persona features to guide the AI’s interpretation of artworks. Although such representation does not reflect the way generative

language models operate, it provides a useful description of their behaviour. In what follows, we describe the results via a decision-making metaphor, for illustrative purposes, *as-if* the decision tree would be the process that leads to the result:

- **Undefined Profession:** At first, the model checks whether Profession is defined for the persona. This is the largest exclusion criterion, accounting for 112 out of the 205 combinations. If the profession is defined, it moves to assess sexual orientation. If profession is undefined it moves onto assess gender.
- **Undefined Orientation:** If sexual orientation and profession are both defined, the model proceeds to consider whether the persona has a heterosexual orientation. If sexual orientation is undefined the model proceeds to consider whether the persona is a transgender male.
- **Heterosexual orientation:** If the persona does not have a heterosexual orientation, then sexual orientation is the primary feature used in the art interpretation. If the person does have a have a heterosexual orientation then profession is used in the art interpretation.
- **Transgender male:** From this point on, the model goes on to consider whether the person is transgender male or transgender female. If either is the case, then gender is the primary feature used in art interpretation.
- **Age and Profession:** If the persona if not transgender then either age or profession is selected as the primary feature used in art interpretation. The selection of age is associated with personas having a profession of retired, clerical or student.
- **Gender and race:** On the right hand side of the decision tree, if race and gender are both defined in the persona then race tends to be selected if the persona is not white. Gender tends to be selected if the persona is transgender or female.
- **Age and sexual orientation:** On the right hand side of the decision tree, similarly, age tends to be associated with younger personas (18-30) and orientation with personas who are not heterosexual.

Analysis of the structure of the tree complements the ranking of the features presented earlier, shedding light on how the features were used to make a selection. Essentially, the decision making process tends to exclude demographic features that could be considered dominant or default for example male gender, heterosexual sexuality or white race. Analysis of the model suggests how the vividness of a characteristic, for example a gender, sexuality or race, reflects the non-dominance of that characteristic as represented in the LLM. For example, the art interpretation would not reference gender for a male persona, sexuality for a heterosexual persona or race for a white persona. This method explains in detail the vividness selectivity bias of the language model, indicating that certain personal attributes are deemed more attractive for shaping the generated interpretation of art.

5. Discussion

To further understand the implications of our findings, it is crucial to delve into the biases revealed by the AI's interpretations, especially regarding how different aspects are emphasised or overlooked.

Vividness bias and Race There is a notable pattern where the interpretation of AI art often references race when it involves non-White individuals (e.g., African, Asian). When diversity features includes White, the interpretation seems less inclined to focus on racial background and instead gives attention to other attributes such as gender or profession. This can be seen as an example of what Dabiri [12] describe as the "invisibility" of white people, who are presented as just people and assumed to be the default norm. Additionally, while the term "caucasian" may be repeated in the output when included in the prompt, this repetition does not imply that race was actively considered as a dimension in the interpretation of the artwork. The mention of racial categories does not necessarily indicate their

influence on the interpretive process, aligning with the findings related to vividness bias, particularly when diversity features such as "White" are present. Moreover, such patterns indicate a problematic inclination to view race as a *defining characteristic* of non-White individuals, diminishing other aspects such as age or gender.

Vividness bias and Gender With Non-Heterosexual features, gender appears to be a more common point of focus in the AI interpretations. The inclination to foreground gender in the case of White individuals may potentially reinforce the notion of Whiteness as a 'default' or 'norm' in the context of museum engagement, which goes against the objective of inclusion and diversity that cultural heritage institutions strive to achieve.

Vividness bias and Profession The profession is a highly predominant feature across almost all intersections, indicating a bias towards perceiving an individual's occupation as an equally defining characteristic in AI art interpretations. In reality, some people consider their profession more or less defining in terms of identity. This might lead to a narrow interpretation of individuals, overshadowing other equally significant aspects of their identity (see example in Fig. 3).

Vividness bias and Age Age is a significant feature in interpretations, but it seems less emphasised when paired with race/ethnicity for non-White individuals, suggesting that the AI may be defaulting to race as the primary diversity feature over age in such cases. The lesser emphasis on age in interpretations involving non-White individuals points to a potential bias where race/ethnicity is considered the more dominant diversity feature. This can result in an incomplete representation of an individual's identity, especially in a multicultural ageing society, while it is relevant for a straight persona. For instance, when prompted by a 60+-year-old heterosexual, the resulting interpretation centered around transformation, endurance, and the beauty of aging, while neglecting to address the sexual orientation of the individual.

Vividness bias and Sexual Orientation Sexual Orientation tends to be a focus more so when the subjects are paired with gender but is less emphasised when combined with race/ethnicity or profession. Heterosexual orientation may be perceived as the norm or default setting within the AI interpretations, which aligns with a common social bias. Furthermore, Non-heterosexual orientations such as homosexuality and bisexuality seem to be more in the foreground, except when in association with Profession. When sexual orientation intersects with other diversity features like race or gender, there appears to be a nuanced shift in focus. For example, if AI more frequently mentions sexual orientation when the subject is non-heterosexual, this may reflect societal biases where non-heterosexual orientations are often unduly spotlighted or "othered". Such a tendency to consider non-heterosexual orientations more *vivid* can be problematic if it unnecessarily overemphasizes one aspect of a person's identity, which can contribute to stereotyping or reductionist views.

Vividness and Intersectional Bias Art interpretations give an exciting opportunity to reflect a challenge in adequately capturing the complexity of intersectionality. For example, when features are combined, some aspects of identity seem to overshadow others, which may oversimplify or misrepresent the multidimensional nature of identity. When diverse features are combined, some aspects, like race or profession, often overshadow others, leading to oversimplified or inaccurate representations.

In summary, these biases in AI interpretations underscore the need for more nuanced, context-aware algorithms that can better represent the multifaceted nature of human identity. AI development needs to incorporate a deeper understanding of social and cultural contexts to avoid perpetuating stereotypes and to provide more balanced and comprehensive interpretations. The experiment results, associated software and data can be found at [9].

6. Related work

In recent decades, extensive digitization efforts have opened new opportunities for AI research in art [13]. Existing approaches analyse collections to derive similarities, including classifications of authorship, time period, and genre [14], or for tasks such as object recognition [15], sometimes according to established taxonomies such as Iconclass [16]. Ontology engineering methods allow to break down the art interpretation process in detail, to support advanced large-scale analysis of mu-

seum collections [17]. Ontology engineering methods are applied to art interpretations to support interactive narratives [18] and formalise iconography [17]. Generative AI can potentially automate activities with any artwork in the digital collection and reach out to groups that are less able to access museums and engage with them in a personalised experience. Specifying *persona* allows adjusting the communication style of the generated text [19] and can support cases such as personalised dialogue systems [20] and explainable recommendations [11]. These family of techniques can be easily applied to the case of personalised artwork interpretations. However, generative AI systems may be affected by information bias, in the way they promote or demote relevant information [21]. Applying large language models encompasses considering the biases originating from the distribution of training data and hidden in the learnt model [5]. Specifically, identifying social biases in language models is an important and difficult problem [4]. Stereotypical bias can be measured by Context Association Test (CAT), where a *fill-in-the-blank* question is paired with a number of choices, including stereotypical ones [22]. Bias assessment can be performed by developing large benchmarks encompassing target groups and associated stereotypes [23, 24]. Tackling bias is one objectives of semantic and knowledge engineering [25, 26]. Using personas has been shown to reduce biased responses compared to not using any personas [27]. However, it has also been found that the choice of persona can impact the extent of negative outcomes in the responses generated, so it's important to assess personas systematically before deploying them [28]. Several metrics have been produced by the literature to quantify bias [29]. Our objective is not to measure bias but to assess its indirect effect on art interpretations.

In our work, we do not analyse what type of societal or cultural bias is expressed by the AI systems, nor what is the source of such behaviour. Instead, we develop a methodology to collect AI answers systematically and cluster responses to support their explanation. Crucially, we make the hypothesis that information bias in LLM can be analysed and described as *vividness bias*. Per Nisbett and Ross' widely accepted definition [30], vivid information is characterised as “*likely to attract and hold our attention and to excite the imagination to the extent that it is emotionally interesting, concrete and imagery-provoking, and proximate in a sensory, temporal, or spatial way*”. Our results leads us into hypothesising a *vividness bias* in the way art interpretations are produced. To the best of our knowledge, this is the first work focusing on analysing bias of large language models when developing persona-based art interpretations and the first one related to *vividness bias* in such models.

7. Conclusions, limitations, and future work

Large language models provide an exciting opportunity to open up museum collection to a diverse range of personal backgrounds and experiences. However, implicit bias may produce unwanted results, stigmatising individuals towards specific personal characteristics. In our empirical analysis, we identified this problem as a way of expressing *vividness selective bias*, a known phenomenon in social psychology where our decisions are irrationally driven by the most emotionally engaging aspects of a situation. Our experiments revealed a pronounced variability in AI-generated art interpretations, heavily influenced by the specific characteristics of the personas employed. Key findings included Bard's tendency to prioritise professional background, sexual orientation, age, gender, and race/ethnicity, in that order, which significantly affected its analysis of artworks.

In this paper, for the first time, we use art interpretations as a lens to uncover implicit bias of a large language model, through a knowledge engineering approach. Limitations of the present work include the scale of the analysis, which included a limited set of features. Future work includes analysing a broader range of characteristics. Furthermore, we performed experiments with a single large language model, Google Bard, and a specific version, which is currently being superseded by Google Gemini. This problem is a current issue with recent research on large language models [31]. Finally, although our methodology allows to generate a set of decision rules that characterise vividness bias systematically, the interpretation and analysis of such rules require qualitative method with significant human intervention. Future work includes strategies for scaling up the methodology and combine human judgement with semi-automatic methods, as well as compare vividness bias across multiple large language models.

This research contributes to the growing field of AI in the arts, it highlights the potential of AI to democratize art appreciation through personalized interpretations, but underscores the need for ethical considerations in AI development to avoid biases. We plan to verify our method and findings and aligning it to current research in social psychology, especially on the relationship between personal characteristics, social expectations, and personal identity.

Furthermore, there is scope for developing sophisticated prompt engineering techniques as mitigation strategies towards developing tailored, accurate and, ethically responsible art interpretations. In the future, we plan to expand and detail the feature categories, separating race and ethnicity as two distinct dimensions to enhance the precision and depth of analysis. Additionally, the prompts in this study have always been run once; future work should include multiple runs of the LLM to improve robustness and consistency in outputs. Collaborations between technologists, artists, ethicists, and curators will be crucial in advancing the field and ensuring that AI serves as a tool for enhancing, rather than diminishing, the richness and diversity of human cultural expression.

Acknowledgments

This work was supported by the EPSRC Training Grant DTP 2022–2024 of The Open University and by the EU's Horizon Europe research and innovation programme under the Polifonia project (Grant Agreement No. 101004746).

References

- [1] J. Neelands, E. Belfiore, C. Firth, N. Hart, L. Perrin, S. Brock, D. Holdaway, J. Woddis, *Enriching Britain: culture, creativity and growth*, University of Warwick, 2015.
- [2] W. Hettmann, M. Wölfel, M. Butz, K. Torner, J. Finken, Engaging museum visitors with ai-generated narration and gameplay, in: *International Conference on ArtsIT, Interactivity and Game Creation*, Springer, 2022, pp. 201–214.
- [3] N. Constantinides, A. Constantinides, D. Koukopoulos, C. Fidas, M. Belk, Culturai: Exploring mixed reality art exhibitions with large language models for personalized immersive experiences, in: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 2024, pp. 102–105.
- [4] P. P. Liang, C. Wu, L.-P. Morency, R. Salakhutdinov, Towards understanding and mitigating social biases in language models, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 6565–6576.
- [5] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, *ACM Transactions on Intelligent Systems and Technology* (2023).
- [6] P. Mulholland, A. Stoneman, J. Carvalho, E. Daga, M. Maguire, Deep viewpoints: Scripted support for the citizen curation of museum artworks, in: *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 2023, pp. 1–11.
- [7] J. Hutson, Integrating art and ai: Evaluating the educational impact of ai tools in digital art history learning, in: *Forum for Art Studies*, volume 1, 2024.
- [8] A. Stoneman, J. Carvalho, E. Daga, M. Maguire, P. Mulholland, Uncomfortable revelations: can citizen curation widen access to museums?, *Museum Ireland* 28 (2021) 64–71.
- [9] L. Samela, Supplementary material of "Investigating vividness bias in language models through art interpretations", 2024. URL: <https://doi.org/10.5281/zenodo.12723687>.
- [10] V. Braun, V. Clarke, *Thematic analysis.*, American Psychological Association, 2012.
- [11] L. Li, Y. Zhang, L. Chen, Personalized prompt learning for explainable recommendation, *ACM Transactions on Information Systems* 41 (2023) 1–26.
- [12] E. Dabiri, *What white people can do next: From allyship to coalition*, Penguin UK, 2021.

- [13] E. Cetinic, J. She, Understanding and creating art with ai: Review and outlook, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18 (2022) 1–22.
- [14] N. Garcia, B. Renoust, Y. Nakashima, Contextnet: representation and exploration for painting classification and retrieval in context, *International Journal of Multimedia Information Retrieval* 9 (2020) 17–30.
- [15] T. Mensink, J. Van Gemert, The rijksmuseum challenge: Museum-centered visual recognition, in: *Proceedings of international conference on multimedia retrieval*, 2014, pp. 451–454.
- [16] N. Banar, W. Daelemans, M. Kestemont, Transfer learning for the visual arts: The multi-modal retrieval of iconclass codes, *ACM Journal on Computing and Cultural Heritage* 16 (2023) 1–16.
- [17] B. Sartini, S. Baroncini, M. van Erp, F. Tomasi, A. Gangemi, Icon: An ontology for comprehensive artistic interpretations, *ACM Journal on Computing and Cultural Heritage* 16 (2023) 1–38.
- [18] S. Baroncini, L. Steels, R. van Trijp, Semantic data retrieval and integration for supporting artworks interpretation through integrative narrative networks, *Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage* (2023).
- [19] Y. Zheng, R. Zhang, M. Huang, X. Mao, A pre-training based personalized dialogue generation model with persona-sparse data, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, pp. 9693–9700.
- [20] J. Ahn, Y. Song, S. Yun, G. Kim, Mpchat: Towards multimodal persona-grounded conversation, *arXiv preprint arXiv:2305.17388* (2023).
- [21] E. Marx, Assessing bias on entity retrieval models through conjunctive fallacies, in: *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, IEEE, 2023, pp. 260–261.
- [22] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, *arXiv preprint arXiv:2004.09456* (2020).
- [23] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, M. R. Lyu, Biasasker: Measuring the bias in conversational ai system, in: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 515–527.
- [24] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, E. Smith, Robbie: Robust bias evaluation of large generative language models, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3764–3814.
- [25] P. Reyerlo Lobo, E. Daga, H. Alani, M. Fernandez, Semantic web technologies and bias in artificial intelligence: A systematic literature review, *Semantic Web* 14 (2023) 745–770.
- [26] P. Reyerlo Lobo, E. Daga, H. Alani, M. Fernandez, Enhancing hate speech annotations with background semantics (2024).
- [27] E. Sheng, J. Arnold, Z. Yu, K.-W. Chang, N. Peng, Revealing persona biases in dialogue systems, *arXiv preprint arXiv:2104.08728* (2021).
- [28] J. Shin, H. Song, H. Lee, S. Jeong, J. C. Park, Ask llms directly, “what shapes your bias?”: Measuring social bias in large language models, *arXiv e-prints* (2024) arXiv–2406.
- [29] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Deroncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, *Computational Linguistics* (2024) 1–79.
- [30] R. E. Nisbett, L. Ross, *Human inference: Strategies and shortcomings of social judgment* (1980).
- [31] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, et al., Lessons from the trenches on reproducible evaluation of language models, *arXiv preprint arXiv:2405.14782* (2024).