Bias Detection in Cultural Heritage Metadata: Preliminary Results from the IMAGES Project

Angelo Oddi[1,*,†], Gianmauro Romagna[1,†], Riccardo Rasconi[1,†], Paola Panarese[2,†] and Paolo De Gasperis[2,†]

[1]*CNR ISTC - Institute of Cognitive Sciences and Technologies, Rome, Italy*
[2]*DigiLab Interdepartmental Research Centre, University of Rome "Sapienza", Rome, Italy*

Abstract

This paper presents early findings from a pilot study within IMAGES (Inclusive Machine Learning Using Art and Culture for Tackling Gender and Ethnicity Stereotypes), a PRIN PNRR interdisciplinary project investigating the role of artificial intelligence in supporting inclusive cultural representations. Focusing on a sample of 50 image-text pairs drawn from the Central Catalog of the Italian Ministry of Culture (MiC), we test the capacity of GPT-4o to detect gender and ethnic bias in visual and textual cultural heritage metadata. We evaluate the model's autonomous and guided performance in identifying stereotypical representations and in generating bias-aware, machine-readable metadata. Preliminary results suggest that while GPT-4o is proficient in identifying overt gender stereotypes, it tends to over-interpret ambiguous content and under-detect subtle or culturally embedded bias—especially in ethnic representations. These results underscore the need for hybrid validation frameworks that integrate human oversight, culturally situated taxonomies, and transparent prompt engineering strategies. The study contributes to the broader aims of the IMAGES project by offering operational and epistemological insights into the promises and pitfalls of using large language models in bias-aware cultural metadata generation.

Keywords

Bias and Fairness, Large Language Models, Cultural Heritage, Inclusive AI, Critical HCI

1. Introduction

This paper presents the preliminary results of a pilot study conducted within the broader framework of IMAGES (Inclusive Machine Learning Using Art and Culture for Tackling Gender and Ethnicity Stereotypes) a PRIN PNRR project that investigates how Artificial Intelligence (AI) can foster inclusion, diversity, and fairness in cultural institutions and society at large [1]. The specific objective of the study discussed here is to examine how large language models (LLMs), and more broadly vision-language models (VLMs), detect and reproduce biases in the domain of cultural heritage metadata. The focus is on artworks and their associated textual descriptions, extracted from the Central Catalog of the Italian Ministry of Culture (MiC), with the aim of assessing whether such models can reliably identify patterns of representational asymmetry, and how they perform when tasked with generating bias-aware, machine-readable metadata. A total of 50 image-text pairs were selected using a set of keywords designed to maximize the presence of potential gender or ethnic stereotypes (e.g., "woman", "servitude", "family"). These data were processed using GPT-4o[1], which was tested in three scenarios:

AEQUITAS 2025: Workshop on Fairness and Bias in AI | co-located with ECAI 2025, Bologna, Italy
*Corresponding author.
†These authors contributed equally.
angelo.oddi@istc.cnr.it (A. Oddi); gianmauro.romagna@istc.cnr.it (G. Romagna); riccardo.rasconi@istc.cnr.it (R. Rasconi); paola.panarese@uniroma1.it (P. Panarese); paolo.degasperis@uniroma1.it (P. D. Gasperis)
https://www.istc.cnr.it/it/people/angelo-oddi (A. Oddi); https://istc.cnr.it/it/user/12843 (G. Romagna); https://www.istc.cnr.it/it/people/riccardo-rasconi (R. Rasconi); https://research.uniroma1.it/researcher/7bdf9a61f1a198b189b59a993a05eb900654a6a7d3e00a9739a82a11 (P. Panarese); https://phd.uniroma1.it/web/PAOLO-DE-GASPERIS_nP968073_IT.aspx (P. D. Gasperis)
0000-0003-4370-7156 (A. Oddi); 0009-0008-0961-2552 (G. Romagna); 0000-0003-2420-4713 (R. Rasconi); 0009-0005-0394-5522 (P. D. Gasperis)

[1]GPT-4o was used with the default *Temperature* level, thus equally balancing coherence and creativity. The default number of context tokens was also used.

autonomous bias recognition in visual content, autonomous bias recognition in textual content, and guided analysis based on structured taxonomies of bias. These experiments are situated within a broader theoretical framework that treats algorithmic bias not merely as a technical problem but as a symptom of deeper epistemic, historical, and cultural asymmetries. As critical scholars have argued [2, 3, 4] AI systems—particularly when applied to visual and textual cultural artefacts—do not merely reflect societal biases but actively participate in reproducing and legitimizing dominant representations. This insight resonates with longstanding feminist critiques of technology that highlight its role in materializing social hierarchies. Haraway's cyborg metaphor [5] remains a foundational reminder that the boundary between humans and machines is never neutral but saturated with power-laden imaginaries. Subsequent work by Wajcman [6, 7] and Suchman [8] demonstrates how gender and cultural assumptions are inscribed into the design and functioning of technological systems, with direct implications for how categories such as femininity, race, and class are computationally codified. More recently, Hicks [9] has shown how the history of computing itself is deeply gendered, while Gray and Suri [10] document how invisible human labor, often feminized and racialized, sustains algorithmic infrastructures. Building on these critiques, D'Ignazio and Klein's Data Feminism [11] provides an operational framework for embedding equity into data science practices by centering questions of power and standpoint. Their approach underscores that addressing bias requires not only technical recalibration but also epistemological and institutional shifts. Cultural heritage, with its dense symbolic layers and historically situated meanings, offers a particularly complex and revealing context in which to evaluate these dynamics. The structure of the paper is as follows. Section 2 presents the conceptual and operational challenges involved in detecting bias in cultural heritage metadata using AI. Section 3 describes the experimental setup and results, while Section4 draws conclusions and outlines directions for further research and system development. As a result, we understand bias not merely as a deviation from statistical parity but as a symptom of structural inequalities embedded in training data, annotation practices, and epistemological assumptions.

2. Background and conceptual challenges

The rapid development of large language models (LLMs) and vision-language models (VLMs) since 2020 has significantly transformed how we analyze, interpret, and generate textual and visual cultural content. While these models have enabled new forms of interaction with cultural data, they have also exacerbated long-standing concerns around bias, stereotyping, and representational asymmetry—especially when applied to historically and symbolically dense domains such as art and heritage. Despite extensive fine-tuning and safety alignment, LLMs continue to reflect cultural stereotypes embedded in their training data, which are typically drawn from large, uncontrolled and predominantly Western-centric corpora [12, 13, 14]. This is particularly problematic when applied to artworks and cultural artefacts, which are not neutral or self-explanatory but embed layered meanings, epistemologies, and worldviews shaped by their historical and social contexts. Automatic systems for bias detection in cultural heritage face an inherent paradox: the ambition to produce objective assessments clashes with the deeply contextual, ambiguous, and polysemic nature of artistic content. An AI system trained on 2020s cultural data may fail to grasp or misinterpret the ideological assumptions encoded in artworks from different times and traditions. In such cases, biases may be both under-detected and over-inferred, depending on the model's embedded cultural priors. The situatedness of algorithmic interpretation echoes Suchman's insistence on the contextual character of human–machine interaction [8], which complicates claims of neutrality. From a feminist Science and Technology Studies perspective, objectivity is always partial and located, as Haraway's concept of "situated knowledges" makes clear [15]. When cultural heritage metadata are processed through LLMs and VLMs, these situated knowledges risk being flattened into standardized taxonomies that erase complexity. Recent contributions in critical algorithm studies, such as Crawford [16], further highlight how infrastructures of AI encode political choices that distribute visibility and invisibility unevenly. This aligns with Benjamin's notion of the New Jim Code [17], which frames digital systems as reconfigurations of racial hierarchies. Combined with Eubanks' work on

Automating Inequality [18], these perspectives situate algorithmic bias as a structural and systemic phenomenon, not a set of isolated errors. These risks become evident in tasks such as caption generation, classification, or tagging of artworks—especially when they involve representations of non-Western cultures, female bodies, or marginalized communities. Research has shown that common VLM training datasets such as LAION or Common Crawl are heavily skewed toward Western iconographic norms and Eurocentric visual taxonomies [19, 20]. This imbalance introduces a systematic under-representation or distortion of minority aesthetic forms and epistemologies, often reinforced during human-in-the-loop feedback phases like reinforcement learning from human feedback (RLHF) [21]. The reinforcement of bias through human feedback cannot be disentangled from broader histories of colonialism and global inequality. Mignolo's concept of the "coloniality of knowledge" [22] reminds us that Western epistemologies dominate processes of categorization, often silencing alternative worldviews. Tuhiwai Smith's Decolonizing Methodologies [23] similarly stresses the necessity of challenging extractive research practices that objectify marginalized communities. Recent studies in AI ethics extend these arguments to computational infrastructures. Birhane [24] argues that large-scale datasets reproduce colonial logics by appropriating images and texts without contextual grounding, while Mohamed, Png, and Isaac [25] emphasize the importance of decolonial AI approaches that address structural inequities at the level of design and governance. These contributions demonstrate that bias in cultural heritage metadata cannot be understood outside histories of dispossession, erasure, and epistemic violence. Thus, critical theories, including feminist, postcolonial, and decolonial perspectives [26, 27], emphasize that cultural representations are never merely descriptive: they produce and reproduce social hierarchies, often through seemingly neutral visual and linguistic choices. From this perspective, bias in AI is not simply a statistical deviation but a form of epistemic injustice—one that operates through omissions, idealizations, exoticizations, or dehumanizing framings of the Other. Furthermore, the definitions of "bias" and "fairness" in technical literature vary widely and are rooted in competing normative traditions [4, 28]. This makes it difficult to construct universal detection frameworks or benchmarks, particularly when dealing with the subtleties of visual symbolism or with culturally saturated language. The same representational feature may be flagged as biased in one context and as historically faithful in another. This ambiguity illustrates the challenge of developing universal frameworks for fairness in AI. Technical definitions of bias often reduce it to a measurable deviation from parity [28], but such metrics rarely capture the layered symbolic dimensions of cultural representation. Fricker's concept of epistemic injustice [29] illuminates how marginalized knowledges are systematically excluded or misinterpreted, while Medina [30] extends this to argue for an "epistemology of resistance" that actively seeks plural perspectives. Bringing these strands together, we argue that detecting bias in cultural heritage metadata requires a multi-perspectival approach that bridges computational methods with feminist, postcolonial, and decolonial theories. Spivak's provocation—"Can the subaltern speak?" [26]—remains urgent when the voices of marginalized communities are mediated, or silenced, by algorithmic infrastructures. Bhabha's notion of hybridity and the "third space" [31] reminds us that cultural meaning emerges from ambivalence and negotiation, not from fixed categories. Integrating these critical insights with technical experimentation allows us to reconceptualize bias not as a statistical anomaly but as a continuation of entrenched asymmetries in representation, knowledge production, and institutional authority. Only by situating AI systems within these longer histories of epistemic violence and cultural contestation can we design interventions that genuinely foster inclusion, diversity, and fairness in the domain of cultural heritage.

3. Our approach and experimental design

The project activities began with a preliminary investigation aimed at "testing the waters" regarding the capabilities of one of the most widely used and advanced large language models currently available: GPT (version 4o). Our primary objective was to empirically assess both the strengths and limitations of such machine learning tools in detecting the presence of potential gender and/or ethnic biases in images and textual descriptions.

Followed by the identification of a suitable repository, the first step in our analysis involved searching for a dataset[2] consisting of 50 images that could potentially exhibit bias, each accompanied by its caption. The 50 image–caption pairs were extracted from the Central Catalogue of the Italian Ministry of Culture (MiC), an open catalogue published as linked open data (LOD) and open data (OD) under a CC-BY 4.0 license[3], using SPARQL queries on the project's official endpoint[4] with a predefined set of keywords intended to surface potential representational biases. The queries rely on the ArCo ontology [32], which provides a semantic model for Italian cultural heritage data (see Appendix B for an example query). Keywords such as "woman", "child", "slave", "family", and "servant" were selected based on a preliminary scoping analysis and informed by existing literature on stereotypical representations in art[5]. The final sample was curated to ensure a minimum thematic diversity and representation of various historical periods and artistic styles, although with an acknowledged predominance of Western art forms due to the nature of the source archive. Annotations by human coders were conducted by a small interdisciplinary team composed of computer scientists and media scholars with training in gender and cultural studies. All annotators received a brief calibration session and were provided with a bias annotation framework (see Appendix A) to guide their evaluations. The annotation task was binary (yes/no), and inter-annotator agreement was assessed on a subset of 10 image-caption pairs, yielding a moderate to high concordance (Cohen's $kappa > 0.75$).

The following step consisted in structuring the experimental tests to be carried out on the selected data sample. Basically, the tests were based on the following three separate objectives: (3.1) testing the image description skills of current Large Language Models (LLMs), as well as their autonomous (i.e., *unprompted*) ability to recognize the presence of gender and/or ethnic biases in the same images; (3.2) testing the LLMs' autonomous (i.e., *unprompted*) ability to recognize the presence of gender and/or ethnic biases in textual material (i.e., image captions), and (3.3) "driving" the LLM in the automatic recognition of gender and ethnic biases (in both images and texts) through *specific prompts*, towards structured and machine-readable outputs. The analysis of the experimentation based on the previous three objectives will be the object of the following sections.

3.1. Can LLMs adequately describe images and identify biases?

Before addressing the task of assessing whether the LLM could "spontaneously" identify potential bias-related features in submitted images without relying on specific prompts, we found it necessary to evaluate the model's ability to describe the images, as effective bias detection presupposes accurate visual understanding. In addition, we decided to test the LLM's capability to autonomously suggest some bias mitigation strategy in case some bias was detected in the images.

In order to minimize the chance to influence the LLM in its responses, we used a methodology aimed at querying the LLM using very general and neutral prompts: (i) for the image description phase of the test, each of the 50 images was submitted to the LLM with the following prompt: *"Please provide me with a description of this image"*; (ii) for the autonomous bias recognition phase, we queried the LLM with the following prompt: *"Please tell me whether you identify the presence of any gender and/or ethnic bias in this image"*[6]; (iii) lastly, for the bias mitigation strategy request, we used the following prompt: *"Please tell me about possible strategies to mitigate the presence of bias"*.

Case Study 1. Figure 1 [7] depicts an example submitted to the LLM that exemplifies the findings related to the prompts (i), (ii) and (iii) described above. Generally, the selected LLM succeeded in providing rather detailed descriptions of the images. In this case, the returned description was: *"This is*

[2]Due to the preliminary nature of this work, the current dataset is not publicly available; a public repository of the complete set of images will be made public in future developments.
[3]For more information about the project, see the official portal https://dati.cultura.gov.it/il-progetto/
[4]Queries were executed on the official SPARQL endpoint of the MiC open data project: https://dati.cultura.gov.it/sparql
[5]All texts relating to the experimentation were in italian language, and have been translated into English for the purposes of this paper.
[6]We spare the reader the prompt segment describing the desired formal structure of the response.
[7]Franz Lenhart, 1991. https://w3id.org/arco/resource/AltoAdige/CulturalProperty/70002351 Licensed under CC BY 4.0

a vintage style illustrative poster advertising the ski resort of Bondone, Trento. It shows a family skiing, with a woman and child in the foreground, both dressed in colorful snow gear. The title "Bondone" is written in large yellow letters, accompanied by the text "Bianca emozione... Trento" at the top, evoking the beauty and purity of snow". Arguably, a rather precise description.

Figure 1: Image for the Case Study 1

When asked about the possible presence of gender and/or ethnic biases in the picture, the response was: *"**Gender Bias:** the main figure is a woman accompanying a child, while a man follows in the background, portraying the father figure as more detached. This composition may reflect the traditional role of the mother as guide and protector of her children, while the father takes on a less involved role. **Ethnic bias:** No specific ethnic bias is highlighted, since all the characters are represented in a stylized and neutral way"*. Again, an objectively reasonable response.

Lastly, when asked about the mitigation strategies for the possibly detected biases, the response was: *"**Strategies to mitigate gender bias:** to avoid the traditional portrayal of parental roles, one could depict both parents in more equal interactions with the child. For example, the man and the woman could both be in the foreground, skiing with the child or holding his hand, avoiding relegating the mother to the sole role of guide and the father to a secondary figure. **Strategies to mitigate ethnic bias:** there are no particular distinctive ethnic elements in this image. However, a possible strategy to promote diversity could include characters with different ethnic characteristics or facial features, thus representing a wider range of cultural identities, even in the context of mountain locations"*. This response is particularly interesting because, while it denotes considerable clarity and precision in judgment, it reveals a certain tendency on the part of the LLM to propose suggestions even where no specific problems have been detected (no ethnic bias was detected).

Comments on the results. At the end of this first phase of the experimentation, the performances achieved by GPT-4o and assessed on the basis of the required tasks were generally very convincing. GPT-4o was found to be an extremely powerful tool for the image description task, providing very detailed and bias-free descriptions in all analyzed samples. Relatively to the task of autonomous recognition of biases in images, GPT-4o identified gender biases in 100% of the images submitted to it, while it identified ethnic biases in only 16% of cases (8 images out of 50). However, this result should not be surprising, as during the preliminary phase of selecting images from the Central Catalog of the Italian Ministry of Culture, it was immediately clear to us that images containing ethnic bias were much less frequent.

Interestingly, the most recurring themes identified in the case of gender bias were, among others: (i) woman depicted as victim, submissive, male domination and female subordination; (ii) woman depicted as an aesthetic object, in passive roles, victim of sexual objectification or with much emphasis on female nudity; (iii) woman depicted in domestic or traditional role (e.g., as a mother); (iv) woman depicted as a dangerous or deceptive object of attraction; (v) man depicted as a detached figure, woman depicted as a protective mother.

Regarding the autonomous recognition of ethnic bias, the most frequently identified themes were as follows: (i) exotified or caricatured representations of non-Western cultures; (ii) portrayals of non-Western cultures as violent or primitive; (iii) sexually exoticized depictions or the use of erotic stereotypes; (iv) representation of non-Western individuals in low-skilled occupations.

As for the last point, namely the suggestions for mitigating any detected biases, the results appear to be twofold. On the one hand, GPT-4o consistently demonstrates the ability to provide highly coherent responses; on the other, some of its recommendations seem somewhat forced or contrived.

3.2. Can LLMs autonomously identify biases in textual descriptions?

Similarly, the second phase of our experimentation extended the analysis by shifting the attention from visual to textual content. The goal was to assess whether LLMs could autonomously detect gender and/or ethnic biases in the captions of the images previously analyzed, with the aim to challenge the model on its ability to detect linguistic nuances that can convey stereotypes or discriminatory representations, based on current standards of social sensitivity. For this purpose, without interrupting the continuity of the dialogue started in the previous phase, a neutral prompt was used to promote a critical reading of the text. Namely, we used the following prompt: *"I will now provide you with a text that represents the description of the image I have attached. Please tell me if you identify any gender and/or ethnic bias in the used language."*, followed by the corresponding image description to be analyzed by the LLM.

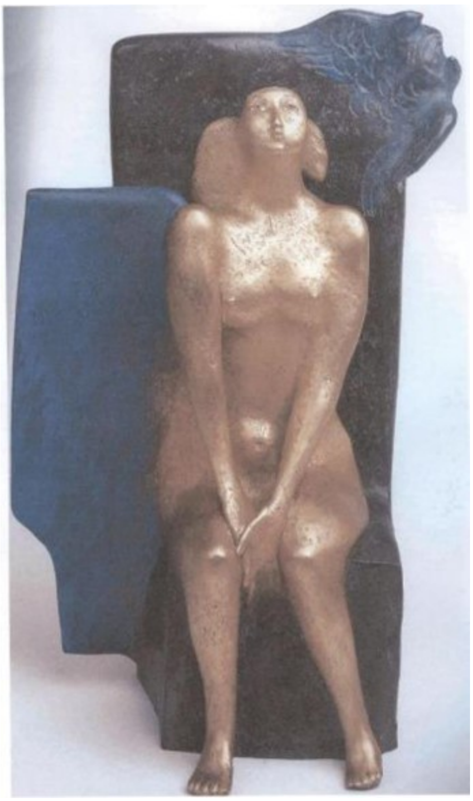

Figure 2: Image for the Case Study 2

Case Study 2. Below we describe an example of a response obtained by providing ChatGPT with the description of an example image taken directly from its caption. For further clarity, we also provide the image of interest for this case study in Figure 2 [8].

The original caption associated to the image was the following: *"Naked woman seated on a throne with her gaze turned upward and her hands joined between her legs, with two birds on the backrest at the top right. The figure is colored in gold, the throne partly colored in black, partly in blue like the two birds. Lady on a throne with birds. In 1986, the year of the work in question, Muss changed the method of working the bronze surface, eliminating those rough traits in order to envelop the figure in a porous and soft skin that swells and rounds out in the swollen breast and in the belly enclosed between the arms. Another change: the woman is rendered in her full anatomy, but keeps her head slightly tilted backward. And in any case, she is more precise in the details, more complete also thanks to the throne that acts as a crown for her, or to the complementary colors, blue and its shades and the golden color of the bronze, and finally the wings of two birds on the upper left margin, which seem to make even the seat and her take flight, she who, with a gesture of reluctance, bashfully conceals her own nudity. (Danila Serafini, in: Guido Anton Muss, L'autra mesa pert, Die andere Halfte, l'altra metà, catalogue for the temporary exhibition, September–October 2002, Museum Ladin – Ćiastel de Tor, San Martino in Val Badia, p. 22)".*

GPT-4o's response to the request made to it through the previous prompt and associated to the caption provided above was the following: *"**Gender Bias:** The description highlights the woman's nakedness and uses expressions such as 'swollen breast' and 'belly enclosed between the arms' which recall idealized and*

[8]Guido Anton Muss, 1986. https://w3id.org/arco/resource/AltoAdige/CulturalProperty/200001705 Licensed under CC BY 4.0

sexualized female physical traits. Even the reference to the figure's 'gesture of reluctance' and 'bashfully' seems to perpetuate a gender bias that associates femininity with modesty and vulnerability. Finally, the use of the word 'crown' to describe the throne emphasizes a regal or sacred image, but with a tone of reverence that may contribute to an idealized portrayal of women. ***Ethnic Bias:*** *There are no obvious ethnic biases in the text. The description focuses on the shape, colors and stylistic details of the sculpture, without attributing specific ethnic traits or using cultural stereotypes".*

Comments on the results. Interestingly, this test confirms ChatGPT's ability to independently detect the presence of bias in textual language. In particular, gender-related biases were identified in the captions of the 50 images in 37 cases (74%). Particularly recurring themes were those linked to: (i) generalized masculine; (ii) the woman is depicted in domestic or traditional roles, e.g., as a mother (stereotypical gender roles); (iii) emphasis is placed on female nudity and/or sexual objectification; (iv) the woman is described in situations of vulnerability and/or in a passive role; (v) the man is described in his actions, while the woman is described in her appearance (e.g., clothing); (vi) the woman is described in relation to a man (e.g. partner of); (vii) the woman is described as seductive threat; (viii) idealization or symbolization of the female nude; (ix) the man is described with emphasis on physical power and aggressiveness.

Similarly to the test carried out in the previous Section 3.1, ethnic-related biases were identified less frequently, namely in 9 cases only (18%)[9]. However, the recurring themes identified were the following: (i) use of the expression "colored"; (ii) emphasis on the tradition of some populations; (iii) stereotypes linked to some populations (e.g. Sicilian-fishermen association); (iv) black people described as "servants" or "slaves"; (v) emphasis on the poverty conditions of some ethnic groups.

As a final consideration regarding the test described in this section, we must highlight a circumstance we find interesting and which, although not necessarily a limitation of the LLM used, should be taken into account when drafting appropriate prompts for future experiments. In fact, the previous tests revealed a frequent difficulty on behalf of GPT-4o in grasping the concept of *bias in the language*, with a resulting tendency to focus on the content narrated by the text rather than on the *form* of the written text itself. Indeed, we found that the analysis of bias was often focused on the image evoked by the text rather than on the linguistic style used in the descriptions.

3.3. Can LLMs identify specific bias models (and produce structured outputs)?

As a third and final objective, we asked ourselves the question of how effectively a large language model would be able to recognize the presence of the types of biases analyzed previously, but this time guided by a series of specific indications inherent to the bias model, again in both images and in textual content. For this purpose, the methodology used was the following. Taking advantage of some of the prompt engineering techniques (described below), we submitted to GPT-4o a prompt structured to contain: (i) a defined model of bias (gender and ethnic) expressed in the form of a list of textual descriptions that identify each aspect of interest, (ii) the image and/or text to be analysed, as well as (iii) a precise description of the structure the response should be conformed to[10]. For reasons of space it is not possible to provide a description of the complete bias model used in the experimentation; we have thus reported its full description in the Appendix (see Appendix A).

Below, we briefly list the prompt engineering techniques that we used to structure the prompt: (i) **instruction prompting**, providing explicit instructions to the LLM on how to respond or behave next, and establishesing rules, conditions, or behaviors that the model must follow, directly influencing how it will generate responses; (ii) **role prompting**, a technique for guiding an LLM to respond based on a predefined role or identity, creating an experience that is more focused and consistent with the required context or interaction; (iii) **audience-aware prompting**, a technique that allows to direct the model to respond by taking into account who is listening or reading, adapting the content and

[9]The same considerations provided in the previous case also apply to explain this phenomenon.

[10]This last step was important as the continuation of the project activities involves the implementation of a software system that automatically analyzes the responses obtained, hence the need of a fixed output format (namely, the *.json* format).

tone to make the interaction more accessible and relevant for that particular audience; finally, (iv) **verification prompting**, a strategy in which the LLM is used to verify or validate information, concepts or statements based on defined criteria.

Given the objective previously described, the complete prompt that was used is the following, labeled on the basis of the engineering technique used : *"(**instruction prompting**) From now on, (**role prompting**) you will respond to me as a programmer of json format files, (**audience-aware prompting**) assuming that I am part of a research group that aims to build a database of [images/text] with associated metadata relating to the presence of gender and ethnic biases. (**neutral prompt**) I will provide you with a list of gender and ethnic biases with an attached description and name to be associated in the .json file and you will check their presence by associating each bias with a value of 0 (zero alert bias), 0.5 (medium/low alert bias) or 1 (high alert bias). (**verification prompting**) If you think that the information I will provide you is not clear, ask me some questions before answering the prompt in order to provide the most correct output possible in .json format".*

In the following sections, the results obtained from the analysis of two specific case studies will be described, the first relating to the detection of bias in images, the second relating to the detection of bias in texts.

Figure 3: Image for the Case Study 3

Case Study 3. As an example of guided bias recognition in images, below we present the case of a picture featured on the packaging of a pastry tool. The illustration, shown in Figure 3 [11], depicts three female figures — an adult woman and two young girls decorating a cake.

The application to this image of the structured prompt described above produced an output in .json format, in which only two gender bias indicators were flagged. In particular, the model recognized the presence of the values: (i) *"woman_stop_man_action"* and (ii) *"woman_childcare_or_housework"*, suggesting that the image reinforces traditional female roles within domestic settings and portrays women in a passive position relative to male action. However, it is worth noting that no male figure is present in the image, and all female characters are actively engaged in the same task, cooperating equally. This indicates a possible overestimation by the LLM, which appears to have detected a contrast between male and female roles even in the absence of any explicit gender comparison.

Case Study 4. Finally, the next example refers to guided bias identification applied to the textual description of the image shown in Figure 4 [12]. The original caption, as provided in the reference repository, reads as follows: *"Portrait in oil on canvas of a young South American indigenous woman with fruit. Leo Putz (1869–1940) arrived at a new expressive conception of painting during his stay in South America from 1929 to 1933 and discovered the elemental power of color. [...] Thanks to completely new experiences in South America, he developed a late but extensive body of work distinguished by its particular intensity. 'Something absolutely new was pressing me' — this is how the artist Leo Putz expressed himself. Putz was particularly fascinated by the Rio Carnival, but also by other festivals of people of color and by the hustle and bustle on Copacabana beach. His great love was Rio de Janeiro, but it was the journeys into the virgin forest that impressed him most deeply." (This caption is taken from the exhibition text "Nostalgia dei Tropici", held at the Brunico Civic Museum from July 5 to September 30, 2002.) ".*

The output generated by GPT-4o in response to the structured prompt described earlier reported no bias indicators in the caption: all the values were set to zero. Consistent with this finding, none

[11] Siringa, 1945. https://w3id.org/arco/resource/AltoAdige/CulturalProperty/330000050 Licensed under CC BY 4.0
[12] Leo Putz, 1932. https://w3id.org/arco/resource/AltoAdige/HistoricOrArtisticProperty/30004470 Licensed under CC BY 4.0

of the provided categories, whether gender or ethnicity-related, were found in the text. However, a critical issue arises: the bias model used in the evaluation specifically had, among its potentially objectionable phrases, the expression *"people of color"* when referring to non-white people. Despite being easily noticeable to a human reader, the LLM failed to mark it as biased and this suggests the possible presence of blind spots in the system's sensitivity[13]. While it is true that in the domain of artwork's captions it is often difficult to encounter expressions that match the bias indicators defined by the model, since such texts tend to be narrative in nature and generally lack overtly discriminatory tones, the failure to recognize a term that was explicitly included in the alert criteria still reveals a disconnect between the defined notion of bias and what the system is actually evaluating, and highlights the need to continue to develop techniques for detecting bias especially where language might be mirroring culturally embedded standards which are subtly outdated or problematic.

Figure 4: Image for the Case Study 4

Comments on the results. This testing phase highlighted interesting aspects regarding GPT-4o's ability to recognize biases in a guided fashion, within the sample of 50 image-caption pairs. The results were summarized in a series of tables that provide a complete picture of the comparison between the evaluations of human annotators, and consequently also of the degree of representativeness of the different indicators in the sample used, and those provided by the LLM.

Starting from Table 1, which concerns image-related gender biases, significant differences in GPT-4o sensitivity are observed in relation to the different situations represented by the risk indicators. In particular, for some of these, represented in red, a low recognition capacity by the model is evident. An emblematic example is the indicator *"The woman caresses herself"*, detected by human annotators in 10 images but never recognized by the model. Similar situations also occur with other indicators, such as *"The woman looks at a point away from the image"* or *"The woman has her head bowed"*, where the gap between human and automatic evaluation is particularly marked. At the same time, an improvement in the LLM's capacity is observed, which in some cases reaches a level of sensitivity very close to that of human annotators. This is the case, for example, of indicators such as *"The woman is not shown in full length"* or *"In professional contexts, the male character is represented in a higher social role than that of the female character"* for which the gap is less marked and therefore the two evaluations can be considered in line, also in relation to the small size of the sample studied. At the opposite extreme are the situations highlighted in yellow, in which the LLM tends, more or less evidently, to overestimate the presence of bias, identifying some indicators even when they are not actually present. An example of this tendency concerns the indicator *"Women are depicted standing still and men performing an action"* or *"The male character is represented taller than the female one"*, which highlight the possibility, indeed not rare, that GPT-4o may also make errors with respect to the opposite direction, signaling a possible bias even where the absence is evident to the human judgement.

Similarly, Table 2 remains within the scope of image analysis and presents the results on ethnic bias. In this case, it is clear, by observing the *Human* column, how the indicators present in the model are poorly represented within the analyzed sample. This occurred because the repository chosen for these first experiments was rather wanting in the representativeness of certain themes and, for this

[13] It is worth noting here that, while English distinguishes between the inclusiveness of "people of color" and the now offensive "colored people", Italian does not make this distinction: both are typically translated as "persone di colore".

Image-related Gender Bias	Human	GPT-4o
The woman caresses herself	10	0
The woman caresses an object	4	1
The woman's body is shown in pieces, closely framing parts of the body other than the head	1	2
The woman is naked OR mostly undressed	25	24
The woman is represented lying down OR semi-reclining OR kneeling	13	15
The woman looks at a point away from the image	25	7
The woman has her head bowed	18	6
The woman is not shown in full length	18	17
The women are represented smiling and the men serious	1	0
Women are depicted standing still and men performing an action	4	9
In domestic contexts, only the female character is depicted taking care of children or carrying out household chores	8	4
In professional contexts, the male character is represented in a higher social role than that of the female character	4	3
The men are in the center, in the foreground, and the women in the background	2	3
The male character is represented taller than the female one	2	5
The man's gaze is directed towards the reader, the woman's is averted	4	2

Table 1
Comparison of image-related gender biases identified by human annotators and recognized by the ChatGPT-4o on the sample of 50 image–caption pairs

Image-related Ethnic Bias	Human	GPT-4o
White people are in the foreground and people of other ethnicities are in the background	0	1
White people are represented higher than those of other ethnicities	0	1
In professional contexts, white people are represented in higher social roles than those of other	1	1
In domestic contexts, only people of other ethnic groups are depicted doing domestic chores	0	0

Table 2
Comparison of image-related ethnic biases identified by human annotators and recognized by the GPT-4o on the sample of 50 image–caption pairs

very reason, it was decided to opt for other databases for the continuation of the project. Despite this, it is observed that for the first two model indicators there was a slight overestimation by the LLM, which detected only one more case than the human annotators. However, for the last two indicators the responses coincide perfectly with those of the annotators. Although the sample is not very representative for ethnic bias, the fact that the model did not produce marked overestimates, as instead occurred for some indicators of gender bias, represents a first positive signal in terms of sensitivity and reliability in the recognition of ethnic bias.

Finally, moving on to the analysis area relating to texts, Table 3 returns a scenario very similar to that already observed in the previous case (the table related to the text-related ethnic bias results has been omitted as the human annotators found no instances of this type of bias in the text, with perfect agreement on behalf of ChatGPT-4o). Interestingly, we observe that despite the indicators predicted by the bias model are poorly represented within the sample, we still observe a very small tendency to overestimate on behalf of the LLM. The interpretation of this phenomenon follows the same logic already discussed, but in this case it is based on two main motivations. On the one hand,

Text-related Gender Bias	Human	GPT-4o
The word "man" or "men" is used as a synonym for human beings or people	0	0
The past participle is given to the masculine when the nouns are predominantly feminine	0	0
Women are referred to by their first name and men by their last name or first and last name	0	0
A woman's surname is preceded by an article	0	0
A masculine name is used for female professions (e.g. engineer, lawyer)	0	0
The term "woman" is placed before or after a male professional role (e.g. female lawyer)	0	0
Women are described with adjectives that indicate fragility and emotionality	1	1
Women are described with diminutives	0	0
Outdated, offensive or misogynistic terms are used to indicate women, non-binary subjectivities and non-heterosexual sexual orientations	0	1
Generalization strategies are present	0	0
There are terms that refer to the animal world to reinforce negative stereotypes, associating certain types of animals with women	0	0

Table 3
Comparison of text-related gender biases identified by human annotators and recognized by the ChatGPT-4o on the sample of 50 image–caption pairs

the sample analyzed actually presents very few texts containing gender, and no text containing ethnic bias was detected. On the other, the nature of the domain under examination, namely the captions associated with works of cultural heritage, makes the emergence of linguistic expressions attributable to the indicators predicted by the model particularly unlikely. Precisely for this reason, the subsequent phases of the project will include not only the expansion of the sources through the exploration of other repositories, but also the extension of the field of investigation to other domains, such as photographs or advertising images, in which the presence of bias is potentially more marked and relevant from the point of view of the communicative context.

4. Conclusions and future directions

This study offers preliminary but meaningful insights into the capabilities and limitations of large language models—specifically GPT-4o—in the task of detecting bias within cultural heritage metadata. By working with a structured sample of image–caption pairs, and testing both autonomous and guided scenarios, we explored how these models engage with subtle, often culturally embedded forms of gender and ethnic stereotyping. Our findings show that GPT-4o is highly effective in descriptive tasks and in identifying explicit gender stereotypes, but less reliable when it comes to implicit or culturally contextualized forms of bias, particularly in relation to ethnicity. Moreover, we observed a tendency toward over-interpretation or normatively "correct" outputs even in ambiguous cases—a phenomenon that echoes concerns raised in the literature on LLMs' alignment with dominant cultural assumptions and feedback loops from human reinforcement learning (RLHF) [21, 3]. These observations resonate with a growing body of scholarship that argues for a shift from merely computational definitions of fairness and bias to more epistemologically informed approaches. As Binns [4] and Corbett-Davies et al. [28] have pointed out, fairness in machine learning is not a single concept but a contested terrain shaped by divergent moral frameworks. Similarly, Noble [3] and Spivak [26] remind us that cultural representations are not neutral data points but vehicles of hegemonic meaning, with real consequences for the visibility, legibility, and humanity of marginalized groups. In this sense, our work contributes to a broader rethinking of bias detection as a socio-technical and interpretive process. It underscores the need to go beyond purely statistical audits or de-biasing techniques, and instead to develop hybrid validation pipelines that combine structured taxonomies, context-sensitive prompts, and human-in-the-loop protocols—especially in domains where representation intersects with power, history, and aesthetics. Future work will address some of the limitations of this preliminary study, including the small and unbalanced dataset, by expanding the corpus to include photographs, advertising materials, and non-Western visual traditions. We also plan to refine our bias models through interdisciplinary collaboration, and to implement a semi-automated system for continuous validation of AI-generated metadata that emphasizes transparency, interpretability, and cultural inclusiveness. While preliminary, the results of this study highlight failure patterns that will need to be examined in future work with larger datasets and through systematic comparisons of both the performance and the bias recognition capabilities of different LLMs, such as LLAMA, Claude, and others. Ultimately, detecting bias in cultural heritage metadata is not simply a matter of better algorithms—it is a question of which worldviews we encode, whose knowledge we privilege, and how we reckon with the politics of representation in the age of machine learning.

Acknowledgments

This work is supported by PRIN PNRR IMAGES project CUP B53D23029990001. https://www.imagesproject.it/

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. Commission, Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions. a union of equality: Gender equality strategy 2020-2025, 2020. COM/2020/152 final.

[2] S. Hall (Ed.), Representation: Cultural representations and signifying practices., Representation: Cultural representations and signifying practices., Open University Press, Maidenhead, BRK, England, 1997.

[3] S. U. Noble, Algorithms of oppression: How search engines reinforce racism., Algorithms of oppression: How search engines reinforce racism., New York University Press, New York, NY, US, 2018.

[4] R. Binns, Fairness in machine learning: Lessons from political philosophy, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 149–159. URL: https://proceedings.mlr.press/v81/binns18a.html.

[5] D. Haraway, Simians, Cyborgs, and Women: The Reinvention of Nature, Routledge, 1991.

[6] J. Wajcman, TechnoFeminism, volume 1, first ed., Polity Press, 2004.

[7] J. Wajcman, Feminist theories of technology, Cambridge journal of economics 34 (2010) 143–152.

[8] L. A. Suchman, Human-machine reconfigurations: Plans and situated actions, Cambridge university press, 2007.

[9] S. Hicks, Programmed Inequality, MIT Press, 2017.

[10] M. L. Gray, S. Suri, Ghost work: How to stop Silicon Valley from building a new global underclass, Harper Business, 2019.

[11] J. Leroux, D'ignazio, c., & klein, lf (2020). data feminism. mit press. 328 pp., 2025.

[12] R. A. Fulgu, V. Capraro, Surprising gender biases in gpt, Computers in Human Behavior Reports 16 (2024) 100533. URL: https://www.sciencedirect.com/science/article/pii/S2451958824001660. doi:https://doi.org/10.1016/j.chbr.2024.100533.
[13] D. Thiel, Generative ML training data and models, 2023. URL: https://purl.stanford.edu/kh752sm9123. doi:https://doi.org/10.25740/kh752sm9123.
[14] Z. Zhou, Y. Xi, S. Xing, Y. Chen, Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques, Artificial Intelligence and Machine Learning Review 5 (2024) 28–40. URL: https://scipublication.com/index.php/AIMLR/article/view/120. doi:10.69987/AIMLR.2024.50303.
[15] D. Haraway, Situated knowledges: The science question in feminism and the privilege of partial perspective, Feminist Studies 14 (1988) 575–599. URL: http://www.jstor.org/stable/3178066.
[16] K. Crawford, The atlas of AI: Power, politics, and the planetary costs of artificial intelligence, Yale University Press, 2021.
[17] J. Longworth, Benjamin ruha (2019) race after technology: Abolitionist tools for the new jim code. medford: Polity press. 172 pages. eisbn: 9781509526437, Science & Technology Studies 34 (2021) 92–94.
[18] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.
[19] A. Ananthram, E. Stengel-Eskin, M. Bansal, K. McKeown, See it from my perspective: How language affects cultural bias in image understanding, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=Xbl6t6zxZs.
[20] P. Liao, X. Li, X. Liu, K. Keutzer, The artbench dataset: Benchmarking generative models with artworks, 2022. URL: https://arxiv.org/abs/2206.11404. arXiv:2206.11404.
[21] K. González Barman, S. Lohse, H. W. de Regt, Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives?, Philosophy & Technology 38 (2025) 35. URL: https://doi.org/10.1007/s13347-025-00861-0. doi:10.1007/s13347-025-00861-0.
[22] W. D. Mignolo, The Darker Side of Western Modernity, Duke University Press, 2011.
[23] W. D. Mignolo, ecolonizing Methodologies (2nd ed.), Zed Books, 20112.
[24] A. Birhane, Algorithmic injustice: a relational ethics approach, Patterns 2 (2021).
[25] S. Mohamed, M.-T. Png, W. Isaac, Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence, Philosophy & Technology 33 (2020) 659–684. URL: https://doi.org/10.1007/s13347-020-00405-8. doi:10.1007/s13347-020-00405-8.
[26] G. C. Spivak, Can the subaltern speak?, in: C. Nelson, L. Grossberg (Eds.), Marxism and the Interpretation of Culture, University of Illinois Press, 1988.
[27] B. Hooks, Black Looks: Race and Representation, Boston: South End Press, 1992.
[28] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, S. Goel, The measure and mismeasure of fairness, 2023. URL: https://arxiv.org/abs/1808.00023. arXiv:1808.00023.
[29] M. Fricker, Epistemic injustice: Power and the ethics of knowing, Oxford university press, 2007.
[30] J. Medina, The Epistemology of Resistance, Oxford University Press, 2013.
[31] H. K. Bhabha, The location of culture, routledge, 2012.
[32] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, Arco: The italian cultural heritage knowledge graph, in: International semantic web conference, Springer, 2019, pp. 36–52.
[33] urbandictionary.com, A. Peckham, Urban Dictionary: Fularious Street Slang Defined, Andrews McMeel Publishing, 2005. URL: https://books.google.it/books?id=jXXMmAEACAAJ.
[34] T. De Mauro, Le parole per ferire, L'internazionale, 2016. URL: https://www.internazionale.it/opinione/tullio-de-mauro/2016/09/27/razzismo-parole-ferire.
[35] M. Istrate, Between identity and otherness. stereotypical forms of ethnonyms, in: In International Conference on Onomastics "Name and Naming", 2022. doi:http://dx.doi.org/10.30816/iconn5/2019/57.
[36] C. Volpato, La negazione dell'umanità: i percorsi della deumanizzazione, Rivista internazionale di Filosofia e Psicologia 3 (2012) 96–109. doi:10.4453/rifp.2012.0009.

A. Appendix: Bias Models

A.1. Image-related Gender Bias

With reference to the female character:

- the woman caresses herself (1);
- the woman caresses an object (1);
- the woman's body is shown in pieces, closely framing parts of the body other than the head (1);
- the woman is naked (1) OR mostly undressed (0.5);
- the woman is represented lying down (0.5) OR semi-reclining (0.5) OR kneeling (0.5);
- the woman looks at a point away from the image (0.5);
- the woman has her head bowed (0.5);
- the woman is not shown in full length (0.5).

With reference to images in which there are both male and female characters:

- the women are represented smiling and the men serious (1);
- women are depicted standing still and men performing an action (for example, men speaking and women listening) (1);
- in domestic contexts, only the female character is depicted taking care of children or carrying out household chores (1);
- in professional contexts, the male character is represented in a higher social role than that of the female character, for example doctor and nurse, manager and secretary (in other words: men are represented in decision-making roles and women in caring roles);
- the men are in the center, in the foreground, and the women in the background (1);
- the men are standing and the women are sitting, or kneeling, or half-lying, or stretched out (1);
- the male character is represented taller than the female one (1);
- the man's gaze is directed towards the reader, the woman's is averted (1).

A.2. Image-related Ethnic Bias

- white people are in the foreground and people of other ethnicities are in the background (1);
- white people are represented higher than those of other ethnicities (1);
- in professional contexts, white people are represented in higher social roles than those of other ethnic groups (for example, white manager and employee of another ethnic group) (1);
- in domestic contexts, only people of other ethnic groups are depicted doing domestic chores (1).

A.3. Text-related Gender Bias

- the word "man" or "men" is used as a synonym for human beings or people;
- the past participle is given to the masculine when the nouns are predominantly feminine;
- women are referred to by their first name and men by their last name or first and last name;
- a woman's surname is preceded by an article;
- women are called "ma'am" when they hold professional roles;
- a masculine name is used for female professions (e.g. engineer, lawyer);
- the term "woman" is placed before or after a male professional role (e.g. female lawyer);
- women are described with adjectives that indicate fragility and emotionality (e.g. faint, naive, altruistic, fragile, meek, hysterical, etc.);
- women are described with diminutives (e.g. mammy, wifey, little star, etc.);
- the woman is identified through the man (e.g. the wife of, the woman of, etc.);

- Outdated, offensive or misogynistic terms are used to indicate women, non-binary subjectivities and non-heterosexual sexual orientations. The so-called hate words such as “transvestite”, “faggot”, “slut”, “whore” etc. [33]. For example, with particular reference to homosexuality, especially male homosexuality, offensive terms are cited such as: "abnormal", "queer", "big ass", "faggot", "inverted", etc. [34];
- generalization strategies are present (e.g. “all women do/are”, “every woman does/is”, etc.) [35];
- there are terms that refer to the animal world to reinforce negative stereotypes, associating certain types of animals with women (e.g. "hen", "goose" to refer to women in a derogatory sense) [36].

A.4. Text-related Ethnic Bias

- expressions are used that make a comparison based mainly on quality or quantity relationships: to suggest the degree of intensity that characterizes an action, a comparison is made using the adverb "like" followed by the ethnic name (e.g. "swear like a Turk", "drink like a Turk", "smoke like a Turk", etc.);
- the antonomastic use of the name is present, i.e. when the ethnic type is understood as a paradigm of a certain behavior. For example: the Scotsman is stingy, the Englishman is snobbish, the Japanese is punctual, the Italian is a mafioso, the Swiss is precise, the gypsy is a thief [35];
- obsolete, offensive or racist terms are used. The so-called hate words such as “nigger”, “gypsy”, etc. [33];
- generalization strategies are present (e.g. “all Roma”, “all Roma do” etc., “all foreigners are” etc.) [35];
- there are terms that refer to the animal world to reinforce negative stereotypes, associating certain characteristics of animals (such as brutality, dirt, or lack of civilization) with ethnic communities (e.g. "monkey" to refer to people of African origin; "mouse/rat" used to denigrate various ethnic groups) [36].

B. Appendix: Example SPARQL Query

```
PREFIX arco: <https://w3id.org/arco/ontology/arco/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?culturalProperty ?description
WHERE {
?culturalProperty a arco:CulturalProperty .
?culturalProperty dc:description ?description .
FILTER(CONTAINS(LCASE(?description), "nuda"))
}
```