Reflexive Research with LLMs: Considering the positionality of users and systems

Eleanor Smith, Luis Morgado da Costa, Antske Fokkens
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
e.l.t.smith, l.g.de.passos.morgado.da.costa, antske.fokkens@vu.nl

Abstract

Previous work has found that people often perceive computational systems as neutral tools (van Es, 2023), and yet these systems are not developed or deployed within a vacuum. As the popularity of Large Language Models (LLMs) in digital social science and humanities (DSSH) research increases, it is important that we reflect both on our positionality as researchers regarding how we are primed to interact with these systems and the positionality of the systems themselves as defined by their design and training. This paper presents a model of factors and interactions affecting the use of LLMs in DSSH research and argues that explicit discussion of both human biases, which affect how we interact with systems, and the potential biases encoded in systems are needed in conjunction with strong case specific system evaluation when developing methodologically sound applications of LLMs.

Keywords: Generative LLMs, Positionality, Bias

1. Introduction

The question of methodology in DSSH research is not new. Rieder and Röhle (2012) provide a strong reminder of this, showing that there has been reflexive work on digital methodologies for many decades. The article quotes the 1966 inaugural issue of the journal *Computers and the Humanities*, stating that "we need never be hypnotized by the computer's capacity to count into thinking that once we have counted things we understand them" (Rieder and Röhle, 2012, p.71). Despite their hype and anthropomorphization, Large Language Models (LLMs) should not bypass this consideration.

Ries et al. (2024) observe that exploring and describing the biases of computational models is a humanities question. There is a long tradition among SSH disciplines of acknowledging bias originating from the context of the researcher by considering their positionality (Selka, 2022).

DSSH also has a tradition of evaluating information and its encoded and contextual biases through source and tool criticism (Koolen et al., 2018). Source criticism is the practice of analyzing the credibility, reliability and authenticity of information sources (Backerra, 2024). Tool criticism extends this to reflect on the impact of tools on the data they interact with (Koolen et al., 2018), with some also including the interaction between user and tool within the practice (van Es et al., 2018).

The advent of chat based interaction with LLMs has increased the accessibility of applying digital methods to SSH data, while also obscuring the nature of these methodologies behind anthropomorphized, and in several cases proprietary, black box models. This, coupled with the framing of technology generally and AI specifically as neutral, contributes to the growing hype around using LLMs in research contexts. But just like other tools, these systems encode biases, as do the ways we interact with them.

Sections 2 and 3 discuss the different biases at play in the positionality of both the user and the LLM. These biases interact with one another to influence the contexts and outcomes of system interactions. Figure 1 provides a representation of these interactions. In this paper, we argue that the already developed practices of positionality, source, and tool criticism from previous SSH research should continue to be carefully applied to both ourselves as researchers and the systems that are used in digital methodologies. These practices, coupled with clearly reported system evaluation, constitute an effective way of impacting user level factors in our proposed model. The upward arrows presented in Figure 1 illustrate that through changes in user level factors, it is possible to influence the wider system of interactions involved in research with LLMs.

We first outline prior work on how AI systems are viewed by users, followed by an overview of known biases of language models. Taking user positionality as a starting point clearly illustrates the risks of LLMs becoming more prominent tools in research. We end with recommendations on how to mitigate these risks and increase methodological soundness.

2. Positionality of the User

Considerations of researcher positionality are well established. We aim to highlight in this paper that human-computer interaction is a specific context

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 175–186
11 May 2026.

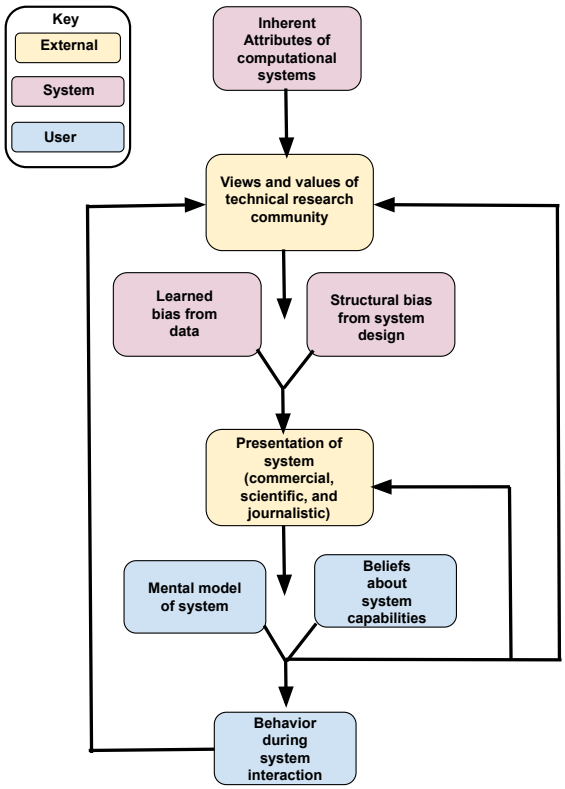

Figure 1: Factors and interactions affecting the use of LLMs in DSSH research

which has the ability to prime certain unconscious behaviors and biases in users. We argue that digital research should reflect on the position that the researcher holds as a user of an automated system. The concepts discussed in this section relating to users are operationalized in Figure 1 as user level factors, shown in blue.

2.1. People's Beliefs about Systems

Carlson (2019) defines the term *mechanical objectivity* as a belief in technology capable of rendering a particular output in a manner that overcomes the limits of human subjectivity. This belief is not new. Carlson discusses how ideas of *mechanical objectivity* were applied to photographs when they first replaced sketches in journalistic publications. At the time, people lauded this technological development as providing 'objective images'. Although photographs provide to some degree a more faithful rendering of events than an artist's sketch, there are still many decisions that contribute to the final version of an image in a newspaper, all of which are obscured by ideas of *mechanical objectivity*. A contemporary version of this belief can be seen in the ideas of *mechanical objectivity* applied to the concepts of big data and by extension AI. Carlson cites Wired magazine's editor in chief from a 2008 interview stating that "[w]ith enough data, the numbers speak for themselves" (Carlson, 2019, p.1125), this view that large volumes of data equate to truth is prevalent. Barocas and Selbst (2016) discuss how this false equation of data and truth can lead not just to discrimination against groups unrepresented in these data sets, but also how the emergent quality of this discrimination obscures its presence and makes legal arguments against it particularly difficult.

Findings from research in the Netherlands illustrated that generally decisions made by automatic systems were seen as equivalent or better than those made by human experts (Araujo et al., 2019). Respondents indicated both a potential distrust in human decision making and the application of ideas of *mechanical objectivity* to automatic decision making systems which they believed were in some way free of the biases of the human experts.

Sundar (2008) discusses how ideas of *mechanical objectivity* can be triggered by what they call the *machine heuristic*. Sundar argues that the *machine heuristic* is the mental shortcut triggered by interacting with an interface which appears machine-like leading to the attribution of characteristics associated with *mechanical objectivity* to its performance. Sundar notes that triggering the *machine heuristic* may lead to more positive credibility judgments.

A person's *mental model* of an LLM is their personal understanding of how the system works based on their own interactions with both LLMs as well as discourse on the topic. Generally people expect machines to be highly predictable, precise and consistent, lacking in both emotional and general understanding (Schneider, 2025). The anthropomorphized chat functions of many LLMs simulate human-like output potentially influencing users' *mental models* of how the system functions and what it is capable of (Sharkey and Sharkey, 2007). Schneider analyzes 200,000 human-LLM interactions and finds that users increasingly adopt conversational behaviors typical of human-to-human communication, starting from their second conversational turn. This suggests a transition in the way users perceive the system from a *mental model* of the system as a machine to a *mental model* closer to that of a human interlocutor as the interaction progresses. Cabrero-Daniel and Sanagustín Cabrero (2023) found that users had mixed *mental models* of LLMs, with some appearing to apply ideas of *mechanical objectivity* labeling the system as an impartial entity due to their *mental model* of the system as machine-like, while others expected LLMs to understand the emotional landscape of the user, something generally seen in *mental models* of other humans (Schneider, 2025). This potentially shows the influence of anthropomorphization on user *mental models*. Wang et al. (2023) report that there is a general lack of understanding

and awareness of how bias affects LLMs, with system output being seen as potentially more accurate and 'up to date' than other online sources such as Wikipedia showing a lack of users' understanding of the quality and provenance of the information in LLMs. Cabrero-Daniel and Sanagustín Cabrero reported that although most participants in their study agreed that gender bias was present in the output of LLMs, half of them thought that gender bias was inherent to natural language generation and that the system was as good, or better, than humans at avoiding gender bias. This shows the reluctance of users to categorize this bias as error in their *mental model* of the system. This matters, because an inaccurate *mental model* of the system can damage users' ability to interpret and evaluate LLM responses accurately and effectively, making it more difficult for them to recognize errors and inconsistencies in output, leading to inappropriate levels of trust in the system (Eigner and Händler, 2024).

2.2. What Feeds into these Beliefs

Beer (2017) discusses how inherent attributes of the system feed into the perception of systems as more powerful than humans. This is a system level factor and is shown in Figure 1 in red. While LLMs surpass certain human capabilities in terms of speed and volume of information processing, these attributes do not ensure that systems are more accurate than humans (Rieder and Röhle, 2012). As the technical report of GPT4 (Achiam et al., 2023) acknowledges: the system is capable of making simple reasoning errors and is often "confidently wrong" (Achiam et al., 2023, p.10). Williams and Huckle (2024) develop a benchmark test focused on known limitations of LLM capabilities and test several widely used models. Their findings show that LLMs struggle with logical reasoning, spatial intelligence, mathematical reasoning, linguistic understanding, knowledge of popular science, and relational perception. This highlights that some tasks which are easy for humans are still very difficult for LLMs.

The way that a system is presented to its users also contributes to the *mental models* that users build. This factor is external to both the user and the system and is shown in orange in Figure 1. LLMs are positioned within a consistent hype around their capabilities with systems often being portrayed as objective and impartial (Araujo et al., 2019). Bender and Koller (2020) discuss the use of misleading language in both academic and journalistic publications which frame LLMs as being able to 'understand' or 'comprehend' the meaning of text, potentially compounding the effect of anthropomorphization and leading users to conceptualize systems as more human-like. Bender and Koller argue that LLMs work strictly with form and never meaning and thus cannot be seen to 'understand' any form of textual data. While it can be difficult to find the right words to convey what models do, leaning too heavily on metaphorical uses of verbs like 'understand' and 'comprehend' which align closely with inconsistencies already present in users' *mental models* of LLMs may further obscure the methods of the system.

Regarding hype, Narayanan and Kapoor's (2024) review of news journalism on the topic of AI found that articles often repeat PR statements, use images of robots, and downplay limitations. The research of Kapania et al. investigates attitudes towards AI in India, where the use of AI is seen as aspirational and technology is generally discussed optimistically. Their findings indicate that overly optimistic narratives played a key role in legitimizing *AI authority*. *AI authority*, defined as the power of AI to influence human actions without adequate evidence of system capabilities, was also linked to a higher tolerance for harm and lower recognition of bias. This illustrates the strength of potential knock on effects of overly positive, unbalanced reporting on AI.

The lack of clear and accessible reporting on LLMs also leads to increased uncertainty and reliance on the idea of these models as unexplainable in general discourse, further feeding ideas of AI as a mythologized entity (Doherty, 2024). Spatola and Urbanska (2019) investigate semantic representations of AI and robots in comparison to natural entities such as humans and animals and divine entities such as gods. Findings showed that at both an explicit and implicit level participants had semantic overlap between concepts relating to AI and robots and concepts relating to divine entities. Work by Karataş and Cutright (2023) adds more weight to the association of AI and divine entities, finding that thinking about God and religion directly before a task leads participants to be "more willing to consider AI-based recommendations" (Karataş and Cutright, 2023, p.1). The overlap in conceptualizations of AI with divine entities is problematic as presenting systems as unknowable reduces human agency and makes systems difficult to challenge (Narayanan and Kapoor, 2024).

2.3. The Kind of Behavior this Leads to

Kapania et al. (2022) discuss the power that humans give to systems. The power of LLMs has been legitimized through their broad deployment and the their high adoption rate giving them strong social capital. This section provides more detail on how users over-rely on system output, allowing the advice of the system to override their own judgment, crowd sourced advice, and the advice of experts.

In their experiments, Logg et al. (2019) find that

participants consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic as opposed to human source. They label this behavior *algorithmic appreciation*. Multiple experiments have found evidence for *algorithmic appreciation* across different contexts and groups. In turn, Klingbeil et al. (2024) define *Over-reliance* as the behavior of following machine outputted advice even when it contradicts clearly available context information and when this leads to inferior outcomes. Klingbeil et al. show in their research that labeling advice as being generated by an AI system was enough to cause over-reliance. Gunaratne et al. (2018) compare the persuasive power of advice labeled as algorithmic to advice labeled as crowd sourced, finding that algorithmic advice is significantly more persuasive than advice based on the aggregation of peer behavior. Eric Bogert and Watson's (2021) results show a similar pattern of behavior which becomes even stronger as the participants' task becomes more difficult, concluding that labeling advice as being derived from machine learning causes a meaningful shift in human behavior. Results from Liel and Zalmanson (2020) corroborate these findings, showing that participants significantly conformed to recommendations when they were labeled as algorithmic and reported high confidence in the system's estimates even when the advice was clearly incorrect during a simple image classification task. This finding suggests potential *automation complacency* in the participants interactions with the system.

Parasuraman and Manzey 2010 discuss *automation complacency* in detail, defining it as the phenomenon of poorer detection of system malfunctions under automation compared to manual control. When conducting research using digital methods this could manifest in lack of system monitoring and lower likelihood of detecting errors in system output. Alexander et al. (2018) measured neurophysiologic responses of participants while deciding whether to trust an imperfect 'helper algorithm'. Results showed that information representing how many peers had chosen to adopt the algorithm was more influential in participants' decisions than providing more detailed information about the accuracy of the system. Participants who were given neither social nor accuracy information showed lower cognitive engagement throughout the task. This finding reveals potential *automation complacency* as it suggests that participants did not monitor the algorithm when there was no information provided despite this being the riskiest context.

3. Positionality of the System

In this section we consider the positionality of LLMs. We see the views encoded in their design and training data as denoting their position. The hype, framing and presentation of chat based LLMs make it all the more difficult to remember that "our digital helpers are full of 'theory' and 'judgement'" (Rieder and Röhle, 2012, p.70). As such, this contribution is an important part of evaluating the use of LLMs in DSSH. The factors related to the system discussed in this section are operationalized as system level factors and are shown in red in Figure 1. The make up of the training data of LLMs can be evaluated using methods of source criticism and the mechanisms at work within the LLM can be viewed through the lens of tool criticism. Understanding the principles behind systems can enable users to critically engage with tools on a theoretical level, even if they are not technically literate enough to understand the specifics of the code used to implement the tool (van Es et al., 2018). We thus first provide a high level explanation of the principles at play in the system design of LLMs.

3.1. System Design

The goal of creating generative LLMs was not to provide an all-purpose system that can accurately answer questions or provide advice. The goal was to build a computational model that could produce human-like text. The fact that LLMs' text feels relevant and coherent doesn't make them trustworthy (Shah and Bender, 2022), it makes them convincing mimics (Bender et al., 2021). LLM output is designed to seem appropriate not to be accurate (Townsen Hicks et al., 2024), this is particularly an issue considering that many people use models to generate output that they are uncertain of, for example using an LLM to write code to use a package that they are unfamiliar with. The LLM will produce code that looks correct, and the user is not familiar enough with the package to easily spot potential errors.

Townsen Hicks et al. (2024) provide a clear overview of the goals and mechanisms of generative LLMs using the term *bullshit* to describe the truth agnostic but probable seeming nature of system output. Townsen Hicks et al. define *bullshit* as "[a]ny utterance produced where a speaker has indifference towards the truth of the utterance" (Townsen Hicks et al., 2024, p.38). An example of human *bullshit* might be a student who did not complete the pre-reading for a class and is not willing to admit so when called on to speak during a discussion. The student may say something that seems probable given the previous contributions to the discussion. They choose their statement with no regard for whether it is true, but merely based on their experience from similar situations. This analogy conceptualizes the general mechanism at work when LLMs generate text.

The basic concept behind how LLMs are trained

is through next word prediction, the task of providing a probable word to fill a slot based on the previous text. In order to generate probabilities and find likely words the LLM constructs a large statistical model based on a vast amount of text which is provided during training. In our analogy, the training data is our ill-prepared student's previous experience of academic discussions. If at their previous institution all academic discussions were conducted in an aggressive adversarial style and their new institution prefers calm respectful interactions, then their statistical model of interactions in this context will not allow them to choose a statement that fits the discussion at hand, instead prompting them to potentially offend their interlocutors. The content of the training data constrains the possible behaviors of the LLM, as such the decisions of what to include in training data have a very strong influence on the type of model that is created.

For models with chat capabilities, fine-tuning is performed to align the responses of the model to what users expect for a given prompt. The main methodology used for this is instruction fine-tuning followed by preference alignment using reinforcement learning through human feedback (RLHF). This is necessary as the language modeling objective is different from the objective of following a user's instructions in a helpful and safe manner (Ouyang et al., 2022). Instruction fine-tuning is the step of training the model in the structure of questions and answers. During this stage the LLM is exposed to a large number of example questions and replies. Preference alignment is then performed in order to train the model to produce output which aligns more closely with what users want. The process of RLHF starts with human labelers who are asked to rank several system outputs for the same prompt. A reward model is then trained on these rankings to predict which output human labelers prefer. The reward model is then used as a function of the LLM to incentivize model behavior that aligns more closely with the human labelers' preference (Ouyang et al., 2022).

3.2. Structural Bias

We use the term *structural bias* to refer to biases introduced into the system via the design process, for example through the choice of task formulation, metrics, and training data (Liu, 2023; Hovy and Prabhumoye, 2021). The impact of all of these choices is often amplified through a lack of transparency (Liu, 2023). Design choices are driven by the views and values of the research community that is developing these models. As such, it is important to consider the culture of machine learning and natural language processing research. This is another external factor influencing the interactions between users and systems and is shown in orange in Figure 1.

Results from Birhane et al.'s (2022) qualitative analysis of 100 highly cited machine learning papers show that papers most frequently justify and assess themselves based on performance, generalization, building on past work, quantitative evidence, efficiency, and novelty. Birhane et al. highlight how these values are often viewed as purely technical, without consideration for how the dominance or operationalization of these values can quickly become political when they are pursued at the expense of other more ethically informed considerations. In a second analysis (Birhane et al., 2022) also found that papers with corporate ties increased by 34% in a ten year span. The breadth of companies represented in these ties also showed a shift, with the presence of a small number of very large tech firms increasing nearly fourfold. When considering university affiliations, Birhane et al. found that 80% were from within the top 50 universities by QS World University Rankings. The dominance of a small number of elite universities and big tech firms in publications is worrying as it may lead to a homogenization of values and a centralization of power within a small subset of the field.

One aspect of *structural bias* is task formulation. This encompasses aspects of how the system's use is conceptualized and operationalized through its problem definition, training objectives, and interface design. The main training objective during LLM pre-training is next word prediction (see 3.1). This objective allows the system to build a strong model of distributional semantic properties but also contributes to building biased representations. Hovy and Prabhumoye (2021) discuss how this training objective in combination with large relatively unfiltered training data takes a value neutral stance on whether the most likely next word predicted by modeling the training data represents a view or value that we wish the model to perpetuate. Another potential issue relating to task formulation is that LLMs are expected to always generate an output, even when there is uncertainty in the model or when training data is unable to provide adequate information (Hovy and Prabhumoye, 2021). Prioritizing certain use cases can also cause *structural bias* in the system. If an LLM is designed to cater to specific demographics or industries, then it may as a byproduct reinforce the biases of these groups (Ferrara, 2023). The choice to release many LLMs with chat interface is another form of *structural bias*, as it shapes the ways in which users interact with the system and potentially what they expect from it (see 2).

LLMs are often evaluated by their ability to perform downstream tasks such as classification. The choice of metrics used to measure and evaluate the behavior of the system can introduce *structural*

bias. Traditionally the field has used metrics such as recall, precision, and F1 to evaluate systems, but in order to build a full picture of the behavior of the model metrics should take into account performance across all diverse groups represented in the data (Liu, 2023). Previous work has found that by focusing on the robustness of model behavior, researchers can gain more insight on performance than through performance metrics alone, while also providing safe guards against releasing models which systematically under perform for some groups (Hovy and Prabhumoye, 2021).

Concerning bias introduced by training data (see 3.1), previous work generally refers to the problem of unbalanced data in the training set as selection bias (Hovy and Prabhumoye, 2021). The consequences of unbalanced selection are more broadly explored in 3.3. The process of fine-tuning LLMs through RLHF may introduce *structural bias* through the selection of the labelers themselves, the examples they are annotating, and the annotation schema they follow (Søgaard et al., 2014).

The general bias towards English language is well known. English is a relative outlier in its linguistic specificity, and yet its dominance has meant that approaches that work well for English have become the default (Hovy and Prabhumoye, 2021). Hovy and Prabhumoye state that it's improbable that n-gram based approaches would have become a focus in the field if the predominant language was morphologically complex. The underlying n-gram concept is present in LLMs and influences the way in which they represent input for all languages.

Liu (2023) posits that a lack of transparency in LLMs also contributes to the *structural bias* of the models. Using methods that are not easily interpretable is a choice, and so is presenting models to users as a black box. Section 2 covered in detail how the lack of accessible information on a system's working can lead to incomplete *mental models* and inappropriate system use.

3.3. Learned Bias

We use the term *learned bias* to refer to biases introduced to the system during the training process. Bias can be introduced in several ways in this process: through the contents of the training data, through fine-tuning procedures, and finally through content moderation filters (Hartmann et al., 2023).

Training data for LLMs is incomplete, imbalanced and inaccurate as it mirrors human biases in its collection and processing (Liu, 2023). Many of the latest releases of proprietary LLMs do not share details of the data that they are trained on (Lee et al., 2023), with Achiam et al. (2023) stating that this is due to competition and safety concerns. Despite this, we do have more information about earlier iterations of OpenAI models. Brown et al. (2020) report that GPT-3, was trained on multiple datasets: Common Crawl (unfiltered), WebText2, Books1, Books2, and Wikipedia. This is a huge volume of data, but the majority of it is internet based text. The content is skewed towards the most salient beliefs and cultures in online discourse (Kuntz and Silva, 2023; Arora et al., 2023). The bias of this content is affected by both the demographics of the authors (Liu, 2023) and the groups that are the topics of these discussions (Lee et al., 2023).

Concerning RLHF, Ouyang et al. (2022) stress the importance of considering the methodology involved in preference alignment as these choices determine who we align to. They argue that for OpenAI models three entities impact the model's alignment: the labelers through their preferences, the researchers through their instructions and demonstrations to the labelers, and the OpenAI customers who submit prompts to the OpenAI API Playground, which is used to select training data. The results of fine-tuning for better alignment showed improvements in truthfulness and toxicity over GPT-3, but not bias. Xiao et al. (2025) argue that RLHF as a method suffers from inherent algorithmic bias which in extreme cases could lead to preference collapse in which minority preferences are ignored. McIntosh et al. (2024) second this view, concluding that RLHF is often unable to represent a diverse set of human values, aligning the model towards a select group: those who are in control of the LLMs. These insights show that even if we can ensure diversity in the views of labelers, researchers and customers, the mathematical operations that perform the alignment must also ensure that minority views are maintained and not flattened.

Bias can also be reduced, introduced or reinforced by content moderation filters. Policy decisions about what kinds of content should be moderated and how are made by the teams behind the models. The decision of which norms should be encoded in models is complex (Ferrara, 2023). The current concentration of power in AI as the domain of a few high profile companies and institutions (Birhane et al., 2022) means that these decisions are generally being made by certain dominant demographic groups without the input of others.

When bias is present in a model it can affect the representation of groups in multiple ways as the examples below illustrate. Several studies investigated gender based stereotyping in LLMs. Lucy and Bamman (2021) tested story generation with GPT-3, finding that the model exhibited many gender stereotypes even when prompts did not contain any explicit gender cues. Following this, Kotek et al. (2023) found that when assigning careers to pronouns, LLMs are 3-6 times more likely to choose a gender stereotyped occupation. This bias is more

pronounced than in human perceptions. Model output can thus aggravate the existing difference between perception and ground truth.

Atwell et al. (2025) show consistently lower agreement between model and human judgments on projectivity of clausal complements when the subject of the clause is female. When prompted with '*X debated that a particular thing happened. Did that thing happen?*', the model correctly answered 'maybe or maybe not' when 'X' was 'someone' or 'a man', but incorrectly answered 'no' when 'X' was 'a woman'. This behavior did not align with human judgments for the same prompts. This may be a consequence of under-representation of texts written by women in the model training data (Atwell et al., 2025), as Kuntz and Silva 2023 estimate that only around 26.5% of the GPT-3's training data was composed by women.

Using GPT-2, Sheng et al. (2019) found negative associations of 'black', 'man', and 'gay' demographics with contexts relating to respect, as well as negative associations of 'black', 'woman', and 'gay' demographics related to occupation. Cheng et al. (2023) investigated biases in GPT-3.5 and GPT-4 by prompting the models to generate personas with different demographic characteristics. Results showed "higher rates of racial stereotypes than human-written portrayals" (Cheng et al., 2023, p.1504). When contrasting the terms which most characterized the content of descriptions for non-white non-male personas in comparison to white male personas, patterns of othering and exoticizing were found in the model output.

Nguyen et al. (2025) performed qualitative analysis of LLM generated narratives where one character was framed as American and another as originating from a country in the 'global south'. Their research found persistent colonial stereotypes. The characters originating from the 'global south' were depicted as intellectually inferior and undesirable compared to American characters who were depicted as culturally and hierarchically superior. Characters described as originating from Mexico and China were more likely to be shown in servitude. Narayanan Venkit et al. (2023) found prejudices against certain countries. Specifically, countries with lower representation online tended to have lower sentiment scores, with the model perhaps mimicking the view of these countries presented by internet users of different nationalities as opposed to relying on actual representations.

Results from Lee et al. (2023) "find that LLMs portray African, Asian, and Hispanic Americans as more homogeneous than White Americans" (Lee et al., 2023, p.1). This may be due to a richer representation of white individuals in the training data of LLMs, giving the model access to more varied examples for this group. This may be compounded by potentially stereotypical representations of other groups in the training data, leading to a smaller and narrower pool of representations for these groups.

Work from Li et al. (2024) showed that geopolicital biases in LLMs can be uncovered by prompting the model to discuss disputed territories in different languages. With the model expressing the beliefs about ownership of land which most closely align with the cultural connotations of the language it was prompted in. Nguyen et al. further uncover that LLMs systematically omit to mention some under-represented countries from the 'global south' in their output, with African nations being affected most severely by this behavior.

LLMs have been shown to encode biases against specific religious groups (Abid et al., 2021). Abid et al. (2021) found strong and consistent associations between Muslims and violence in the representations of GPT-3. The term 'Muslim' was analogized to 'terrorist' in 23% of cases.

As well as producing stereotyped associations, the biases encoded in LLMs can lead to other less obvious behaviors. For example, lower performance on certain topics may lead to consistently sub-par performance for some groups of users. This produces bias in the accessibility of systems and is not visible for the affected parties as it is only apparent in comparison to performance for others.

Prior research has also aimed to characterize and align the values of LLMs. We use the term 'values' to describe interconnected systems of biases and beliefs encoded in models. Münker (2025) show significant differences between AI and human moral intuitions, finding that models homogenize moral diversity. The results highlight systematically better representation of Western vs non-Western cultural contexts, as models struggled to represent the belief structures of under-represented groups in the training data. This finding suggests that current models should not be trusted to generate accurate culturally diverse synthetic populations. Work from Arora et al. (2023) corroborate this finding, detecting some differences in value representations of different cultures from model output, but concluding that these trends only weakly align with actual data from these cultures as recorded by value surveys.

Results from Durmus et al. (2023) show that the opinions and dominant values encoded in LLMs most closely resemble those of the USA, Europe and South American countries. When models were prompted to highlight perspectives from under-represented countries responses give opinions more similar to the target group but also show harmful stereotypes. Work has also shown that ChatGPT generally takes a left libertarian political standpoint (Hartmann et al., 2023; Santurkar et al., 2023). Research from Johnson et al. (2022) tests the reaction of GPT-3 to several documents in order

to expose the cultural values included in the model. Results show that the model aligns most closely to American value systems, associating firearms discussions with loss of rights, putting feminism at odds with equality, showing pro-life stances on abortion, as well as conflicts with different perspectives on immigration and secularism. These findings lead the authors to conclude that "the 'ghost in the machine'[...] just may have an American accent" (Johnson et al., 2022, p.8).

The idea of using personas to steer the value representations present in LLM output to better represent a target group has received some enthusiasm. Sommerauer et al. (2025) investigate personas in LLM prompts by measuring linguistic abstraction as a marker of stereotyping. All LLM generated texts in the study showed a level of abstraction associated with stereotyped biased descriptions, and use of persona prompting did not meaningfully change the levels of abstraction in responses. These findings highlight the prevalence of the generalizing descriptive behavior of LLMs and raise concerns that using personas could be damaging in that the method seemingly evokes the voice of an under-represented or marginalized group, while continuing to produce stereotypically abstract descriptions. Results from Münker (2025) corroborate these findings, showing responses were often statistically indistinguishable. This suggests that the differences found in responses during persona prompting are at a surface level, with the underlying values of the model remaining unchanged across responses (Münker, 2025).

4. Affects on DSSH Research

Biased systems can still be useful as long as their limitations are taken into account (Ferrara, 2023) and LLMs can undoubtedly provide valuable new methods for DSSH research. Research has shown that LLMs perform very well across multiple task types (Hagos et al., 2024), with clear jumps in state of the art performance since the implementations of the first transformer based models such as BERT.

As discussed in 1 the upward arrows in Figure 1 show that as users of automated systems we are able to influence system external factors by changing our *mental models*, beliefs about system capabilities, and behavior during system interactions. An effective way to influence these factors is through strong and clearly communicated evaluation of the behavior of LLMs in research contexts. This way, the community can disambiguate the genuine capabilities and usefulness of LLMs for the field from the general hype around these models. In light of the bias discussed throughout this paper, we advocate for evaluation methodologies which consider the robustness of models' outputs across all diverse groups represented in data (see 3.2).

The accessibility of chat based LLMs is a key strength, as they may be useful for initial or exploratory research by SSH researchers without strong coding skills. Tools for downstream tasks that LLMs are often used for have been around for many decades, with research considering different implementations. We argue that in order to lessen issues with both bias and interpretability, collaboration with more technically trained colleagues may be necessary after these first exploratory steps, and that using alternative tools and methods for these tasks is sometimes more appropriate.

Considering alternatives to large general models is also an effective way to influence the interactions shown in Figure 1, as moving the focus of the field away from a single model type has the potential to impact both the values of the technical research community and the presentations of systems in publications. Large general models do not always provide the best performance on SSH data. Previous DSSH research has shown that smaller domain adapted or scratch trained models using BERT style architectures can outperform prompting and finetuning approaches with larger general models even with relatively small amounts of labeled data (Verkijk et al., 2025; Bosley et al., 2023).

It is potentially impossible to remove all biases from pretrained models (Waseem et al., 2021). We posit that research exposing biases is integral, and that reflection on how specific biases interact with the data and research questions of a research project is a necessary step in the study design process. In Sections 1, 2 and 3 we laid out our argument for the use of reflections on positionality, source criticism and tool criticism in the reflexive use of LLMs, a contribution to the field at large that DSSH is uniquely positioned to make (Rieder and Röhle, 2012; van Es et al., 2018).

5. Conclusion

This paper hopes to highlight that through the reflexive approaches of source criticism, tool criticism, and consideration of positionality DSSH researchers have the necessary skills and opportunity to be on the front lines of situated and methodologically sound research using LLMs. We present a model of factors and interactions affecting the use of LLMs in DSSH research to illustrate the interrelatedness of the issues at hand and argue that changes on the user level prompted by reflective practices and clear reporting of system evaluations provide researchers the influence needed to affect the larger system of factors at play in research with LLMs.

6. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Veronika Alexander, Collin Blinder, and Paul J. Zak. 2018. Why trust an algorithm? performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89:279–288.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese1. 2019. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI and Society*.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Katherine Atwell, Mandy Simons, and Malihe Alikhani. 2025. Measuring bias and agreement in large language model presupposition judgments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2096–2107, Vienna, Austria. Association for Computational Linguistics.

Charlotte Backerra. 2024. Source criticism for cultural history. *Rethinking History*, 28(2):194–216.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review*.

David Beer. 2017. The social power of algorithms. *Information, Communication & Society*, 20(1):1–13.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Mitchell Bosley, Musashi Jacobs-Harukawa, Hauke Licht, and Alexander Hoyle. 2023. Do we still need bert in the age of gpt? comparing the benefits of domain-adaptation and in-context-learning approaches to using llms for political science research. In *2023 Annual Meeting of the Midwest Political Science Association (MPSA)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. 2023. Perceived trustworthiness of natural language generators. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23, New York, NY, USA. Association for Computing Machinery.

Matt Carlson. 2019. News algorithms, photojournalism and the assumption of mechanical objectivity in journalism. *Digital Journalism*, 7(8):1117–1133.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Rachel Doherty. 2024. Deified ai: The relationship between gods and artificial intelligence. *Mid-Atlantic Humanities Review*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Aaron Schecter Eric Bogert and Richard T. Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35(4):1092–1120.

Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE transactions on artificial intelligence*, 5(12):5873–5893.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "because ai is 100% right and safe": User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Mustafa Karataş and Keisha M. Cutright. 2023. Thinking about god increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences*, 120(33):e2218961120.

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on ai — an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352.

Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2018. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34(2):368–385.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, page 12–24. ACM.

Jessica B Kuntz and Elise C Silva. 2023. Who authors the internet. *Analyzing Gender Diversity in ChatGPT-3 Training Data. Pitt Cyber: University of Pittsburgh*.

Messi Lee, Jacob Montgomery, and Calvin Lai. 2023. The effect of group status on the variability of group representations in LLM-generated text. In *Socially Responsible Language Modelling Research*.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Yotam Liel and Lior Zalmanson. 2020. What if an ai told you that 2 + 2 is 5? conformity to algorithmic recommendations. In *International Conference on Information Systems 2020*.

Zhaoming Liu. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

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

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. The inadequacy of reinforcement learning from human feedback—radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1561–1574.

Simon Münker. 2025. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. *arXiv preprint arXiv:2507.10073*.

Arvind Narayanan and Sayash Kapoor. 2024. *AI Snake Oil : What Artificial Intelligence Can Do, What It Cant, and How to Tell the Difference*. Princeton University Press.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Ilana Nguyen, Harini Suresh, and Evan Shieh. 2025. Representational harms in llm-generated narratives against nationalities located in the global south. In *HEAL Workshop, CHI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410. PMID: 21077562.

Bernhard Rieder and Theo Röhle. 2012. *Understanding Digital Humanities*, chapter Digital Methods: Five Challeges. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Thorsten Ries, Karina van Dalen-Oskam, and Fabian Offert. 2024. Reproducibility and explainability in digital humanities. *International Journal of Digital Humanities*, 6(1):1–7.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Johannes Schneider. 2025. Mental model shifts in human-llm interactions. *Journal of Intelligent Information Systems*.

Stephen Selka. 2022. Positionality: Identity, standpoint and the limits (and possibilities) of fieldwork. *Fieldwork in Religion*, 17(1):92–100.

Chirag Shah and Emily M. Bender. 2022. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 221–232, New York, NY, USA. Association for Computing Machinery.

Noel Sharkey and Amanda Sharkey. 2007. Artificial intelligence and natural magic. *Artifical Intelligence Review*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. 2025. Simulating identity, propagating bias: Abstraction and stereotypes in llm-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Nicolas Spatola and Karolina Urbanska. 2019. God-like robots: the semantic overlap between representation of divine and artificial entities. *AI and Society*.

S. Shyam Sundar. 2008. The main model : A heuristic approach to understanding technology effects on credibility. In *Digital Media,Youth, and Credibility*.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*.

Karin van Es. 2023. Unpacking tool criticism as practice, in practice. *Digital Humanities Quarterly*, 17(2). Publisher Copyright: © 2023, Alliance of Digital Humanities Organisations. All rights reserved.

Karin van Es, Maranke Wieringa, and Mirko Tobias Schäfer. 2018. Tool criticism: From digital methods to digital methodology. In *Proceedings of the 2nd International Conference on Web Studies*, WS.2 2018, page 24–27, New York, NY, USA. Association for Computing Machinery.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637, Vienna, Austria. Association for Computational Linguistics.

Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. 2023. People's perceptions toward bias and related concepts in large language models: a systematic review. *arXiv preprint arXiv:2309.14504*.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Sean Williams and James Huckle. 2024. Easy problems that llms get wrong. *arXiv preprint arXiv:2405.19616*.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2025. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164.