Reporting Guidelines for Large Language Models in Human-Robot Interaction

CYNTHIA MATUSZEK, University of Maryland, Baltimore County, USA
TOM WILLIAMS, Colorado School of Mines, USA
NICK DEPALMA, Semio Community, USA
ROSS MEAD, Semio AI, Inc., USA
RUCHEN WEN, University of Maryland, Baltimore County, USA
EIKE SCHNEIDERS, University of Southampton, United Kingdom
CASEY KENNINGTON, Boise State University, USA
ALEMITU BEZABIH, Colorado School of Mines, USA

The comparatively recent advent of Large Language Models (LLMs) has resulted in a wide array of new capabilities and components relevant to Human-Robot Interaction (HRI) researchers. LLMs are being applied to vision, manipulation, planning, reasoning, learning, and HRI problems, frequently as "Scarecrows," in which LLMs serve as black box modules integrated into robot architectures for the purpose of quickly enabling full-pipeline solutions. However, despite this explosion of applications, general questions remain about the best ways to incorporate LLMs into robot architectures, appropriate safety and guardrail considerations, and, critically, how to report properly on HRI research that involves LLMs. In this article, we explore the question of reporting guidelines for HRI researchers who utilize Scarecrows in robot architectures. We identify five key stakeholder groups in the HRI research process, discuss what information each group needs from HRI researchers, and identify appropriate mechanisms for conveying that information from HRI researchers to stakeholders either directly or indirectly. We contribute a set of suggested guidelines regarding what information should be included when researchers disseminate information about HRI research that uses LLMs.

CCS Concepts: • **Computer systems organization → Robotics**; • **Human-centered computing → Natural language interfaces**.

Additional Key Words and Phrases: Scarecrows, Large Language Models, Human-Robot Interaction, Reporting Guidelines, Ethics

ACM Reference Format:
Cynthia Matuszek, Tom Williams, Nick DePalma, Ross Mead, Ruchen Wen, Eike Schneiders, Casey Kennington, and Alemitu Bezabih. 2025. Reporting Guidelines for Large Language Models in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* X, X, Article X (X 2025), 24 pages. https://doi.org/XX.XXXX/XXXXXXX

Authors' Contact Information: Cynthia Matuszek, cmat@umbc.edu, University of Maryland, Baltimore County, Baltimore, Maryland, USA; Tom Williams, twilliams@mines.edu, Colorado School of Mines, Golden, Colorado, USA; Nick DePalma, ndepalma@alum.mit.edu, Semio Community, Los Angeles, California, USA; Ross Mead, ross@semio.ai, Semio AI, Inc., Los Angeles, California, USA; Ruchen Wen, rwen@umbc.edu, University of Maryland, Baltimore County, Baltimore, Maryland, USA; Eike Schneiders, eike.schneiders@soton.ac.uk, University of Southampton, Southampton, United Kingdom; Casey Kennington, caseykennington@boisestate.edu, Boise State University, Boise, Idaho, USA; Alemitu Bezabih, alemitubezabih@mines.edu, Colorado School of Mines, Golden, Colorado, USA.

ACM 2573-9522/2025/X-ARTX
https://doi.org/XX.XXXX/XXXXXXX

1 Introduction

Large Language Models (LLMs) and vision-language models (VLMs) have the power to reshape how we think about interactive robotic systems, allowing for new capabilities and interactions that were previously unreachable. LLMs are being applied to supporting human language interactions [48] and human-robot teaming [55], providing novel methods for robotic planning [19, 89, 102] and decision-making [62], and allowing new approaches to longstanding problems, such as object recognition [128] and task and motion planning [26], *inter alia*. However, the use of LLMs may also constrain the ways that Human-Robot Interaction (HRI) researchers go about designing interactive robots, the types of interactions they choose to design, and the types of scientific questions they choose to investigate. Moreover, LLMs present a wide range of ethical risks that must be safeguarded against and made transparent when they are used in robot design [117].

While many of these considerations are endemic to the general use of LLMs, the use of LLMs in robot architectures comes with unique challenges, especially given that that LLMs are often used as constituent components of robot architectures, responsible for piecemeal tasks as well as for enabling end-to-end human-robot dialogue. For example, recent work has described how LLMs may be used as *Scarecrows*: "'brainless,' straw-man black box modules integrated into robot architectures for the purpose of quickly enabling full-pipeline solutions, much like the use of 'Wizard of Oz' (WoZ) and other human-in-the-loop approaches" [118].

Identifying this parallel between LLMs and WoZ techniques helps to illuminate key responsibilities borne by HRI researchers who choose to use LLMs in their robot architectures. Just as the use of WoZ gives designers the responsibility to report on exactly where, why, and how they use WoZ in the name of transparency and replicability [91], so too does the use of LLMs give designers reporting responsibilities. In this paper, we explore these reporting needs, building both on (a) our previous perspective piece discussing the risks and challenges associated with the use of LLM Scarecrows in HRI [118], and (b) the results of a workshop held at the 2024 ACM/IEEE International Conference on Human-Robot Interaction (organized by the first five authors and attended by all authors) at which preliminary reporting guidelines were discussed.

A key difference in reporting needs relative to WoZ is the range of **stakeholders** to whom information about the use of LLMs must be disclosed (see Figure 1). Discussions about transparency regarding the use of WoZ techniques have been largely focused on transparency *to other researchers*, due to the typical use of WoZ as a design and experimental methodology rather than a technical tool (excepting possible, arguably questionable uses of WoZ-like techniques that fall under the umbrella of "Artificial Artificial Intelligence" [86]). However, because LLMs are typically used (or envisioned to be used) on robots deployed in the world and not just constrained to the laboratory, a wider array of reporting considerations must be made with respect to a broader array of stakeholders who will interact with robots, enable interactions with robots, or constrain interactions with robots.

In this paper, we identify (1) a set of stakeholder groups to whom researchers owe a reporting duty, (2) the reasons why each of these types of stakeholders needs information about the use of LLMs, and (3) the information and mode(s) of presentation best aligned with the reporting needs of each of those stakeholder groups. These were developed based on the workshop described above. At this session, invited speakers presented background information about Scarecrows and LLM use, after which attendees separated into groups and brainstormed entries into a worksheet with stakeholders, data to be reported, and the reasoning behind those requirements. These worksheets were shared with attendees and discussed, then synthesized into a consensus table, which the authors elaborated on based on reporting literature (see table 1). We close with suggested reporting guidelines for HRI researchers who are incorporating LLMs into their architectures.

2 Related Work

LLMs have the potential to revolutionize researchers' approach to HRI, and it has not taken long for the research community to take advantage of this opportunity. Due to overwhelming interest in the topic, the focus of our literature review covers work specifically on models and architectures that are used in direct or indirect interaction with end users.

Using an LLM as a module in a larger architecture is a common approach to the embodied use of LLMs. Izquierdo-Badiola et al. [42] present a system using an LLM as a module in collaboration with end users that generates joint plans to be executed with end users. Sun et al. [105] use an LLM to perform collaborative task planning by interacting directly with a Partially Observable Markov Decision Process (POMDP). Nwankwo and Rueckert [84] present a collaborative robot that uses both an LLM and a VLM to capture situational awareness for the rest of the robot architecture.

Likewise, understanding user instructions is a common use case for LLMs. Wu et al. [121, 123] demonstrate the effectiveness of LLMs as modules for taking instructions and generating effective behaviors for users in a variety of tasks (e.g., tidying up, or selecting and moving objects based on instructions). Mahadevan et al. [69] extend previous work in LLMs to accept feedback when generating code that synthesizes expressive robot behaviors.

Other researchers have used LLMs to handle robot dialogue [1], or to reason in a shallow way about emotional context [75] and social norms [130]. Finally, some researchers have shown specific limitations of LLMs in HRI tasks, such as theory of mind [111] and moral reasoning [92]. Due to more general limitations of LLMs, some researchers have begun to analyze the contexts in which LLMs should or should not be used [49]. Authors often express the risks to end users by using LLMs in their architecture, typically found in their discussions (e.g., [92, 111, 130, 131]). Yet other related areas like stakeholder requirements remain unexplored; for example, none of the papers we reviewed mention the provenance of the data on which models are trained (cf. [121]). Thus, there is a significant need for an analysis of the risks of LLMs in HRI and when and how those risks should be reported.

One key question to consider is when the reporting guidelines discussed in this paper are specific to the use of Large Language Models in HRI, as opposed to being more broadly applicable to machine learning in the field, and fact there is significant overlap. While there are many methodologies to studying and investigating human-robot interaction, one popular perspective is to use data-driven approaches to generate interactive, social, safe robotic behavior [46, 78, 98]. These approaches typically use machine learning to learn policies, behaviors, and speech patterns from sensor data, animations, and other useful data. However, machine learning is a large field and has many different practices, some explainable and some not. Historically, machine learning has used a number of approaches to learn a useful, human understandable *representation*. Some of these representations are stochastic in nature [80], while others are inspired by natural neural networks [38] and require expert inspection to understand how connections impact network output. The breakthrough work that we are discussing in this paper involves using end-to-end neural networks that use transformers applied to language without explicitly modeling linguistic phenomena in a more piecewise fashion [87].

Historically, models in data-driven approaches have involved piecewise learning and optimization of component parts of an overall behavior architecture. While these pieces are compatible in a way that allows them to be simultaneously learned, they can be broken apart and inspected for improvements and feature development. However, when we discuss explainability in transformer-based neural networks, we are addressing the challenge of understanding the connections between embeddings that were learned and how they impact overall model outcomes. Focusing on what a transformer has 'learned' has proven challenging to make guarantees about. These types of models

are somewhat black box in nature, making it hard to generalize what phenomena these models have learned [64, 134]. While some progress has been made, the lack of guarantees inherent to these transformer-based models significantly affects the safety and explainability of these models [67], leading to a need for clear, explicit reporting on how and whether they are to be used. When we discuss the challenge and lack of explainability in these models, we refer to not only in providing guarantees but to the models' opacity [95]. The apparent linguistic capabilities of LLMs make them particularly tempting to incorporate in human-facing applications in robotics; it is these transformer-based, language-capable models we specifically refer to in this paper.

3 The "Who?": Stakeholder Categories

We begin by identifying five key groups of stakeholders to whom researchers incorporating LLMs might owe reporting duties, either directly on indirectly, depending on the use case. Although HRI researchers are not the only group that provides information—for example, practitioners who design robot architectures and build robots will likely provide specification sheets to decision makers—in this article *we focus on information originating from the HRI research process.*

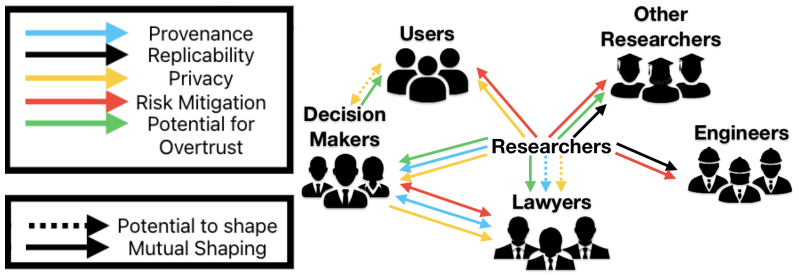

Fig. 1. Human-robot interaction researchers using large language models have a responsibility to communicate critical information to a variety of stakeholders (Section 3). Each stakeholder group is interested in different types of information, including: Data and model *provenance*, or sourcing; whether a system can be reliably *replicated* and appropriate information to support such replication; how and whether a system infringes on *user privacy* and how that infringement can be managed; what other *risks* result from using an LLM in an HRI system, and how to mitigate those risks; and whether and how such systems may lead to user *overtrust*, along with suggestions for mitigation. Different stakeholder groups have the ability to shape (affect and report on) these axes and the information that supports reasoning about them.

We briefly define each stakeholder group before discussing their reporting needs in the following sections. Figure 1 shows some of the types of information that should be shared between and among groups of stakeholders.

Researchers in HRI are increasingly using, extending, building, and/or designing LLMs to meet their research goals. These researchers may determine the possible inputs, outputs, and underlying model architecture to answer a specific research question. For example, research questions can be specific to a model training regime or model architecture, and are typically sensitive to characteristics–such as the transparency and bias in model outputs, key model limitations, model hardware requirements, model power consumption and environmental impacts–as well as the larger economic and societal impacts of LLMs. Moreover, researchers and other members of the broader research community may need to evaluate the ethical risks of LLMs and their use in robot architectures as part of their interactions with Institutional Review Boards (IRBs).

Stakeholder	Why do they need reporting?	What should be reported, and how?
Researchers	Replicability, Long-term risk assessment	Technical information on architecture, models, datasets, prompts, and ethical risks, presented in research paper sections.
Engineers	Verification, Medium-term risk assessment	Model dependencies, inputs, and outputs, known failure modes and ethical risks, presented in technical reports and summarized in model cards.
End Users	Near-term risk assessment (to self and known others)	Influence of LLMs on system behavior and known ethical risks, presented through quick-start guides and tutorials, education and tech literacy programs, and transparent interactions.
Decision Makers	Near- and medium-term risk assessment (to self and unknown others)	Influence of LLMs on system behaviors and known ethical risks, presented through model cards, public information/specification sheets, and formal training and certification programs.
Lawyers and Regulators	Community-level risk assessment and response	Ethical risks and legal considerations, presented through policy briefs and publicly accessible documentation.

Table 1. Condensed reporting needs for LLMs in Human-Robot Interaction for five major stakeholder groups. These entries were developed by the authors based on input from the 30+ workshop attendees at the 2024 Conference on Human-Robot Interaction (see section 1 for details).

Engineers build or use existing LLMs, often for a specific use case. Engineers are not usually the stakeholders that make changes to the models themselves; rather, engineers are interested in the surrounding technologies necessary for model access, which could mean a software development access point or a chat interface. Engineers and practitioners are responsible for implementing and scaling the ideas and capabilities developed by researchers, and depend directly on reporting by those researchers.

End Users make use of LLM-enabled technologies. End users are less concerned with how the models work and more concerned with whether a model is useful and what risks it has. End users include home users, hobbyists, students, and professionals such as K–12 educators.[1] In a professional setting, this group of stakeholders may not have complete volition over whether to use provided tools (e.g., nurses and medication delivery robots). End users depend on information provided by researchers both directly (through general education efforts and popular reporting on current science) and indirectly (through implementations provided by engineers).

Decision Makers are those who choose to acquire a robot for use by themselves or others, whether in a home environment (e.g., parents acquiring a robot for their children) or a professional setting (e.g., hospital or school administrators). It is worth noting that this group may have significant overlap with both end users and with other professional units (e.g., approval/acquisitions managers). Decision makers depend partly on data from researchers filtered through engineers, by way of public data, specification sheets, and marketing material.

Lawyers and Regulators are concerned with legal and societal impacts of LLMs, including data used for training (e.g., fair-use or copyrighted data), end-user impacts, environmental impacts, economic impacts, safety and security, as well as sourcing data.[2] Because this group takes a higher level view, their dependence on information comes directly from researchers, from engineers and practitioners, and from population-level education and publicity efforts.

[1]In this work, we exclude people in the vicinity of LLM-using robots who are not direct users, for example, bystanders near a museum kiosk robot. While such users have potentially relevant concerns (e.g., privacy), they fall outside the scope of our present discussion.

[2]For example, Resolution 604 adopted by the American Bar Association: https://www.americanbar.org/content/dam/aba/directories/policy/midyear-2023/604-midyear-2023.pdf

These groups are presented in Table 1, along with a summary of the information that should be reported to each group, and some suggestions as to what reporting media might be appropriate to each group. We elaborate upon these suggestions in the following section.

4 The "Why?" and The "What?": Transparency and Types of Information to be Provided

Having delineated five key groups of stakeholders, we consider why each group needs transparency in the use of LLMs in robot architectures for HRI. For some groups, this may be related to the unique goals of the group, for which transparency is needed to adequately fulfill. For other groups, this may be the unique sources of risk to which that group is vulnerable, for which transparency is necessary to avoid. Figure 1 shows some possible areas of concern that may need to be addressed by researchers and various stakeholders.

Having considered why stakeholders might need transparency into the use of LLMs in robot architectures, we also make suggestions as to what types of information would help each stakeholder group to meet their unique goals or avoid their unique sources of risk, and how those types of information should be presented.

4.1 Researchers

Researchers must provide clarity about their use of LLMs for three key reasons: (1) to improve replicability and scientific integrity, (2) to avoid overtrust, and (3) to ensure an understanding of the costs associated with the research.

To assess the rigor and maturity of the research presented, transparency is critical. Therefore, to understand the context in which work was carried out as well as to replicate the findings, key information needs to be provided by the researchers to other researchers. This includes, for example, the prompts used, the model version number used, the date at which the evaluation was carried out, and information about parameters and context; for a suggested list of what details should be provided, see Section 7. However, given the non-deterministic nature of LLMs, we acknowledge that a replication producing the same results might not be guaranteed even when details are provided.

Beyond strengthening and advancing the quality of research, transparency is vital to minimize the risks that arise from robot overtrust. By understanding the limitations and capabilities of a system and the research conducted on/with it, risks to the integrity of the research community can be minimized, as well as the risk to potential future users of the developed systems.

Lastly, working with LLMs—whether via a browser, a smartphone, an API, or a robot—is a costly endeavor. Estimates have been made that indicate that the slightest performance improvement to translation tasks costs around $150,000 in additional training costs [7]. By disclosing information on the complexity of the trained model (e.g., what data was used or how many parameters), it becomes possible to estimate financial and environmental costs and impacts.

In addition to the need for clarity in reporting on their own research, researchers and members of the broader research community need to fulfill roles as members of Institutional Review Boards (IRBs). Because IRBs are charged with providing ethical and regulatory oversight of research involving human subjects, IRB members need to have a comprehensive understanding of what LLMs are involved in the research or development processes, how those LLMs are used, and what potential impacts this might have on research participants. This understanding can help IRB members to establish and refine LLM-related institutional regulations to prevent negative outcomes (as well as misconduct) during the experimentation or development process.

How to provide transparency among researchers. Researchers using LLMs as part of robot architectures need to disclose a variety of information to ensure replicability of their work:

(1) First and foremost, researchers should disclose the larger architectural context within which LLMs were used, and the specific locations and temporal dynamics of use of those LLMs.
(2) Because the behavior of LLMs is largely determined both by the data on which LLMs were trained and tested, as well as by the prompts provided to them, researchers should also report at a minimum (1) information available (e.g., relevant datasheets [34]) about the data on which their chosen LLM was trained, as well as the nature and origin of any supplemental data the researchers used for fine-tuning; and (2) both the exact prompts used for testing, and the history of prompts that preceded those testing queries.
(3) Because the capabilities and limitations of LLMs are constantly changing, researchers should report not only the version number of the LLM(s) they use, but also the date on which queries were given to those LLMs, and the LLM seed specified, if any. (Although additional reliability measures are desirable, these are frequently difficult with black box systems; authors should consider additional methods for capturing system performance under different parameters.)
(4) Researchers who use LLMs in their overall system should also report the range of inputs, as well as the modifications to the outputs that allow those outputs to be useful for the downstream users of the LLMs.
(5) Researchers reporting in scientific papers on their use of LLMs should be clear about the known ethical risks of those LLMs, including privacy, economic, environmental, and social justice concerns, and how those risks were/were not mitigated through the researchers' technical approach. In doing so, researchers should be clear not only about the types of risks raised by their use of LLMs, but also the likelihood, severity, and timescale of those risks.

While the risks associated with LLMs can impact anyone, research participants in human-robot interaction studies involving LLMs can be uniquely vulnerable because they are the first ones exposed to these technologies while they are under development; planning to report on the above points will help researchers explicitly consider such risks. This information should all be reported in research papers (distributed as appropriate across "Technical Approach" and "Discussion" sections of those papers), and reported in IRB applications. For example, the following details might be presented in a paper's "Technical Approach" section:

> *"All evaluations performed in this work leveraged OpenAI's gpt-4-0613, a text-only LLM with a 8,192 token context window. All queries were given to this model between 04/01/2024 and 04/30/2024. No seed was specified, and a temperature of 0.5 was used. One-shot prompts were used, with no prior chat conversation provided. Each natural language instruction verbally provided by the user to the Softbank Pepper (described in Section X) in the context of interaction (described in Section Y) was translated from speech to text using the PocketSphinx 5.0.3 Automatic Speech Recognition Model, and then fed to the LLM, with the prompt:*
>
> I am going to give you a sentence. You will return the logical representation of the utterance, with no explanation or anything beyond the logical representation. *[Here, the authors would provide the examples they are using to enable in-context learning, which we have omitted for readability.]*
>
> Here is the sentence: "Put the box on the large table in the kitchen".
>
> *Results returned by the LLM were then provided to the natural language understanding system as described in Section Z."*

The following text might similarly appear in the paper's "Discussion" section and/or as part of the research project's initial IRB application.

> The use of LLMs comes with a number of privacy, economic, environmental, and social justice concerns that may be propagated through the use of LLM-based methods. Our use of GPT4 to parse interactants' utterances means that those utterances are implicitly provided to OpenAI; a risk with high likelihood

but low severity, that will occur once per interactant utterance. Other researchers hoping to use this technique should make this privacy risk transparent to users, request a Zero Data Retention endpoint from OpenAI, or use an LLM that does not present this privacy risk.

Our use of GPT4 also presents several economic risks, due to the water and electricity consumption of LLMs, and the reliance of LLMs and other AI systems on toxic mining practices. Each GPT4 query is estimated to consume 0.3kWh, and at least as much water (17ml) as GPT3. These high energy costs mean that this type of model may be unsuited to domains where a high rate of interaction is required over long periods of time.

Finally, the use of GPT4 presents social justice concerns. Some of these concerns, such as the tendency of LLMs to center and normalize White, Western, masculine voices and perspectives (a manifestation of roboticists' power in the cultural domain [115]), are mitigated through our limited use of GPT4 to only effect relatively straightforward transformations of text into logical representation. Other concerns, like the reliance of OpenAI's ethical guardrails on practices of data colonialism, where Kenyan workers are subjected to psychologically and economically exploited conditions to benefit American tech companies, are not similarly mitigated. While further damage is not inflicted upon repeated uses of GPT4 within our architecture, this nevertheless presents hidden costs that may limit the ability of our model to ethically be transitioned into use cases beyond the laboratory, unless a different LLM is used.

This text explains where LLMs are used, how the use of LLMs might affect outcomes, what risks LLMs present to users or experimental participants, how those risks are bounded by the way they are used in the robot architecture. When used in the context of an IRB application, researchers should supplement this type of explanation with a further explanation of when and how these risks would be disclosed to experimental participants.

4.2 Engineers

Engineers and robotics practitioners require insight into the capabilities, failure modes, and mechanisms of LLM-based systems for several reasons. First and foremost, in designing and building robots for deployment, they must decide whether, how, and how frequently to use large models and systems that depend on them, with an emphasis on building systems that perform safely and as intended. As a matter of responsible engineering conduct, professional engineers must build systems that do not behave unethically or support unethical behavior. They also require transparency into the use of LLMs in robot architectures in part because they have associated reporting burdens to their own stakeholders, and in part to support understanding and debugging their own systems. Ultimately, engineers are responsible for the overall design, verification, performance, and maintenance of a robotic system—tasks they cannot perform without clear information from researchers who design the underlying models.

How to provide transparency to engineers. The use of an LLM provides many of the same opportunities [11, 39] and challenges [81] that come with scaling experimental systems for deployment. Understanding the risks and limitations of LLMs will allow engineers to provide guarantees that an LLM does not introduce critical flaws to a robot architecture that would prevent safe deployment to the end user. This is important both for good engineering practice, and so that engineers can report costs and risks to their own stakeholders. Engineers will also need to report known flaws in the system and, where possible, remediate these flaws from occurring, in part via careful prompt engineering [99]. Prompt engineering has been a popular method of introducing LLMs into a much larger robot architecture. Using LLMs as components requires extensive knowledge of fault and failure cases in prompt engineering, such that the details used in researcher experiments provide valuable insights to preventing these flaws from occurring in the first place. It is worth noting that, while "guardrails" are a major resource for engineers to manage risks and flaws [33, 79, 113, 129],

current guardrail technology is insufficient for deployed systems that incorporate LLMs [20]. This insufficiency can be observed through LLM's continued tendency to engage in sexist or otherwise hateful speech [37, 124], and to engage in 'bullshitting' [119, cf.] and other actions that present risks to democracy [22]. Moreover, guardrails on their own are insufficient to address the cultural and ethnic homogenization of LLM-driven robot speech [18] (although cp. [93]), and in some cases may even reinforce such homogenization (due to the ways that rule-driven guardrail systems inherit from racist philosophies that equate morality with White standards of behavior [74]) [116, p. 121–124]. Finally, because guardrails are often enabled using Reinforcement Learning from Human Feedback, they naturally encode the biases of those providing feedback, meaning that current guardrail-based ethic systems will likely encode and reinforce White supremacist and patriarchal behavioral expectations [116, p. 87].

Depending on the role and function of an LLM within a robot architecture, engineers may treat the LLM as a black box with a particular set of expected inputs and outputs, or they may be more concerned with *how* the LLM was trained or what the model itself may learn as they may be applying the experiment to their own data. In fact, Jiang et al. [44] found that the top concerns for machine learning engineers when reusing models are data provenance, reproducibility, and portability. This reflects the broad spectrum of use cases and responsibilities of engineers in this context and provides barriers and challenges to reporting this information to engineers.

To communicate meaningful knowledge from researchers to engineers is no easy task, as this detailed information can be quite verbose. Appropriate details include hyperparameters, data origin, including copyright and licensing information, as well as the the model's propensity to hallucinate. It is also important that engineers know what the model has learned, what kinds of outputs are expected from the model, and in what contexts it is appropriate for it to run efficiently and effectively. Recently, new ways of concisely communicating summaries of complex information have emerged that support this need, such as model cards [73, 76], or "nutrition facts" style visualizations [36, 110], such as that shown in Figure 2. These types of visualizations incorporate other social and reproducibility considerations, such as why the data was used, with what hyperparameters, and what model details should be discussed.

AI Nutrition Facts

LaPSe

Description An LLM-Based Parsing System used in the CRA Cognitive Robotic Arhictecture	
Privacy Ladder Level	N/A
Feature is Optional	No
Model Type	Generative
Base Model	OpenAI - GPT-4
Trust Ingredients	
Base Model Trained with Customer Data No offline training beyond GPT-4 Base Model. WARNING: GPT-4 Training may have involved exploitative practices [1]	N/A
Customer Data is Shared with Model Vendor	Yes
Training Data Anonymized	No
Data Deletion	No
Human in the Loop	No
Data Retention	Data Provided to OpenAI
Compliance	
Logging & Auditing	No
Guardrails Unlikely to be a concern given nature of task	No
Input/Output Consistency Expected Input: natural language utterances such as "Go to the room at the end of the long hallway" Expected output: Logical Representation such as "Command(speaker, self, go-to(self,X)),[room(X), hallway(Y), long(Y), at-end(X,Y)]" comprised of an Utterance, speaker, hearer, base semantics, and supplemental semantics. Accuracy of 95% achieved in informal tests.	Yes
Other Resources Energy Use: ~0.3kWh /Utterance, ~17ml H20 /Utterance [1] https://time.com/6247678 /openai-chatgpt-kenya-workers/	

Learn more about this label at **nutrition-facts.ai**

Fig. 2. Sample Nutrition Label, generated through https://nutrition-facts.ai.

While these mechanisms for communicating information can often provide a helpful resource about model details and training regimens, they may be limited in their ability to communicate more scientific components of trained models. For instance, how models are used may typically not be found on model cards but are important details for engineers. Many times, training parameters are only available via paper reporting or via websites, such as Hugging Face or Neptune. On the other hand, risks, user evaluations, and model transparency details are frequently still lacking from these sources. For this reason, it is still recommended that

researchers provide as much supplementary materials as is possible and necessary to address the needs of engineers.

4.3 End Users

End users are not a monolithic group, and HRI researchers need to consider the possible characteristics of different user groups when providing information (further discussed in Section 6.4). As an example, some populations do not have an *obligation* to interact with robotic systems, such as most users of social companion robots, who are in a position to determine when to opt out of two-way communication or to opt out of use entirely. These decisions can only be made in the presence of sufficient information about the system's behavior and data flow patterns. The needs of this group may vary from, for example, those of professional end users, for whom robots exist in workplaces and serve professional functions (e.g., nurses who interact with medication-dispensing robots in hospitals). For this group, technology use may not be a matter of choice, changing the nature of their interaction with those technologies [10]. As such, these users must be provided with enough information to understand the nature of the interactions, allowing them to understand possible risks and avoid overtrust/overreliance on the technology [52].

Some users may also play the role of a mediator of technology—for example, a patient in a hospital receiving medication from a robot, whose introduction to and questions about the technology will likely be handled by the nurses who work with it. In this situation, a clear understanding of the robot technology will affect technological understanding and appropriate use patterns [21] by these secondary users. All end users require a clear understanding of the actual capabilities of the robot and to what degree these systems should be trusted (described below), as either under- or over-trust can lead to system misuse [52].

How to provide transparency to end users. Users frequently have relatively little control over and minimal visibility into the nature of the technology underlying robots and related technologies (e.g., voice-activated assistants [43]). Accordingly, information that should be made available to end users is frequently distinct from the information engineers and researchers need to exchange. This information can be difficult to formalize, because "end users" is a heterogeneous category whose members may require different kinds of information. Nonetheless, providing the following information should be considered a baseline.

First, researchers and engineers should make it clear to users when interactions are generated in part by an LLM, using general terminology and readily understandable examples. This should be paired with disclaimers that the system may not be reliable, and that any information or suggestions made by the system may be incorrect. Because LLMs are frequently used for back-end tasks as well as direct interaction, this will likely involve providing samples of possible failure modes. It may also be useful to convey the concept of "guardrails" with explanations of their weaknesses (although this may convey information about how to subvert those safety measures).

Users should also be made aware of the potential privacy failures and data leakage associated with use of online models or other advanced features, such as personalized agents. Specifically, personalization mechanisms deployed on LLMs often rely on retaining users' conversation history and historical behaviors [109, 133]. These data pipelines should be explained, making clear what data will be retained, who may have access to user-supplied data, and how users can opt out of those features. Conveying privacy information successfully is nontrivial [9, 35, 53] and may not be desirable to the providers of technology [132], meaning that this is both an important requirement and difficult to actualize. Developers of user-facing technologies should therefore be aware of best practices in providing privacy explanations ([12, 13, 100], *inter alia*), and avoid "dark patterns" in presenting information (e.g., presenting risk information only in jargon-filled license agreements).

Similarly, users should be made aware of mechanisms for and associated implications of declining the use of LLM-based technology.

These types of information could be provided to users in at least two forms. First, some information could be provided to users who are the "first line" of robot deployment through quick-start guides and tutorials integrated into the unboxing and setup process. These guides and tutorials could provide an point of intervention where users can be sure to be exposed to key information about the design and limitations of robots' software architectures before they are able to use those robots. Second, some information could be provided by robots themselves through modalities, such as natural language, either during introductions or over the course of interactions. In the same way that ChatGPT might caveat some dialogue turns with a statement of its limitations as a language model, LLM-enabled robots might find opportunities to indicate where and how LLMs are used as part of relevant interactions.

4.4 Decision Makers

While many end users may have little decision making power as to when, where, and how robots are deployed in their contexts of use, some users and non-users will have decision making capability. We define *decision makers* as those with the power to decide to purchase and deploy LLM-enabled robots; as such, this group requires a clear understanding of these systems not only to determine *how* to use robots in a variety of settings, but *whether or not* to do so—deciding whether to buy robots, whether to deploy them, how they should be used, and how frequently to activate them. This group relies on researchers, technology vendors, and others to make information broadly available so that decisions can be made in an informed way. "Decision makers" here loosely incorporates or overlaps with such professional groups as acquirers, approvers, managers, lab leaders, and administrators, but also non-professionals, such as parents.

This group requires information about how LLMs affect overall performance of the system under consideration (e.g., what capabilities are enabled, and also what risks are introduced), but must also consider their significant ethical implications [40, 107]. Of particular note are the privacy implications of deploying LLM-using robots; concerns about the privacy implications of modern smart automation [29, 71] already represent an (often well-founded) barrier to adoption [41, 97], and physical agents with sensors have the potential to compound both real and perceived privacy risks. Additionally, LLMs have been demonstrated to produce social dependency behaviors [85, 125] and can lead to over-reliance among some users [57], which may be exacerbated by the human tendency to over-trust physically embodied agents [2, 112].

It is also worth noting that decision makers may be responsible for making decisions *for others*—including vulnerable groups (e.g., children, whose parents may make decisions regarding robots in their environment), or groups that do not have a choice about whether to use robots (e.g., when tool use is required in a workplace setting). The ability of these stakeholders to make decisions about the appropriateness of robot systems at a broader scale implies the need for even more transparency with respect to the pros, cons, risks, and rewards of those systems.

How to provide transparency to decision makers. Many of the types of information that should be provided to those who will make decisions regarding the purchase and deployment of LLM-enabled robots are similar to those that should be provided to end users; however, this group has additional needs and requires different mechanisms for conveying information. Notably, it is necessary to convey information about the specific risks posed *before* decisions are made about when and whether to acquire and deploy these tools; this includes providing information about not only possible failure modes (e.g., acting from a biased perspective, hallucinating factual information, and contributing to harmful outputs), but also about possible longer-term costs (e.g., cultural homogenization and

environmental impacts [107]). Decision makers may also have a significant need to more deeply understand the possibility of data leakage, in which queries and interactions are fed back into the system as training data, making it possible both to retrieve those interactions [6, 14, 16, 135] and to draw inferences about the querent and their environment [103].

Again, this information should be conveyed clearly and in a fashion tailored to the specific use case, using examples as appropriate. However, to make informed decisions, one must weigh risks and trade-offs when choosing whether and how to use a system. Many of the mechanisms for informing users prior to system use (e.g., unboxing videos, introductory tutorials, and reviews) do not present risks and trade-offs in a meaningful manner. Like engineers, this population may benefit from reporting in the form of model cards, and, like end users, public information efforts are key. However, they additionally require the existence of public information, such as clear, easily interpreted product information and specification sheets, made available via mechanisms such as web sites. Groups or companies that provide LLM-enabled robots should make every effort to make such information publicly available.

In addition to general community awareness programs, decision-makers in professional settings may also benefit from more formal training and certification programs. While such programs would not provide sufficient transparency about individual models, they could provide professionals with awareness of the types of risks and considerations they should be attentive towards when considering the use of particular models.

4.5 Lawyers and Regulators

Lawyers, policy makers, and regulators must design, enact, and enforce laws, guidelines, and specifications to promote the responsible, accountable, and sustainable design of robotic systems. HRI researchers should support this group in obtaining a comprehensive but less technical understanding of the processes involved in designing, developing, deploying, and using physical systems in day-to-day life, including in situations where those systems incorporate LLMs. Policy makers must be aware of how LLMs in such systems have the potential to adversely affect individuals, communities, society, and the environment, so that they can make effective decisions.

There are multiple axes of transparency regarding what information policy makers should be aware of (at some level of abstraction). LLMs are varied as well as complex, but policy makers should be consistently be made aware of *at least* the sourcing of training data, the collection and utilization of user data and associated privacy implications, and the probability and implications of hallucinations and other failures. The source and usage of training data for large models is important because it determines the behavior of those models; however, for policy makers they are particularly important because there are significant controversies surrounding the ownership of that data—numerous lawsuits have been filed, and different jurisdictions have varying rules regarding the use of publicly available data for training [3]. Many of the largest providers of LLM models, such as OpenAI and Google, do not reveal exactly what data their models are trained on, and it is difficult to reverse engineer whether a particular source is included [70], arguably exacerbating the need for oversight and regulation.

Policy makers also need to be aware of how LLMs interact with user data and the associated extensive privacy implications [63]. Large models hosted by developers almost universally take user inputs as additional training data, despite the fact that these models can be induced to repeat training data [15, 17] and can infer surprisingly detailed personal information from interactions [103]. Regulatory privacy frameworks such as the European General Data Protection Regulation (GDPR) and the European AI Act struggle to address opaque, closed-source systems, which frequently use data in ways not conceived of by these laws [30, 56, 65]. These and related concerns necessitate a holistic, rigorous regulation mechanism in the development of LLM-backed robotic applications.

For policy makers, clear information should emphasize the unique challenges posed by LLM-powered robotic systems, which may arise due to the embodiment and physical presence of the robots delivering LLM-related output [106, 122]. These challenges include, but are not limited to, the *epistemological*, the *ecological*, and the *economic*.

Epistemologically, LLMs can affect society by shaping knowledge practices and peoples' understanding of themselves and their societies. Associated threats include the spread of misinformation and disinformation [5, 83], as LLMs in physical systems are in a uniquely persuasive position [66] to disseminate incorrect or false information with or without human intent, as well as generate biased content that favors specific viewpoints [27, 127].

Ecologically, LLM-based robots have real consequences on the environment and the planet at large. The substantial computational resources required to train and operate these robots contribute to skyrocketing energy consumption [104], and the consumption of cold freshwater for cooling large data centers [61].

Economically, data cleaning and annotation crowd work involves a vast network of people, often with unfair compensation rates [45]. Privacy concerns also arise from the collection and use of personal data in LLM training datasets without explicit consent or awareness [120, 126]; these concerns are exacerbated by the physical presence of robots in human spaces [117].

How to provide transparency to policy makers. Overall, the risks policy makers must consider stem from the combination of the data, algorithms, and physical presence of the robot in human spaces. These risks are partially associated with challenges in ensuring the quality and appropriateness of training data and algorithms used to fine-tune LLMs—which imply the need for data and algorithmic regulations [32, 54, 101]—and partially result from the fact that language-using physical agents pose additional risks [117]. This is particularly true given peoples' tendency to overestimate systems that seem capable in a particular area [47].

HRI researchers should therefore work toward transparently informing regulatory stakeholders about the risks associated with robotics platforms that rely on language and, in particular, on large pre-trained models. This includes providing information about the nature and whereabouts of data used for training [34]; for example, regulators might need to know whether an LLM-backed robot uses data from the Internet, which might not be accurate, confidential, free, or representative, and whether the data curation of an LLM involves fair treatment of data cleaners or annotators behind the scenes [82]. They also need to understand some of the risks associated with physical systems in terms of their influence over people, privacy implications, and effects on perceived identity (e.g., perpetuating gender stereotypes [96]), and to be informed of possible immediate and downstream failure modes of such robots.

In addition to supporting the regulation process, the Artificial Intelligence (AI) research community should engage with supporting the regulation of algorithms that are used to train LLMs and apply them to physically situated problems. While regulatory bodies do not need to understand the sophisticated implementations of algorithms, they should be informed of overall mechanisms and how LLMs can go wrong when embedded into robot architectures. Researchers might also consider creating tools or frameworks for algorithm and behavior inspection [88].

This information could be reported in the form of policy briefs and publicly accessible documentation. The Colorado AI Act [23], for example, requires developers to provide (1) disclosures and documentation of model risks and data use to deployers; (2) statements of the purpose, intended use, and means of operation to consumers; and (3) statements regarding high-risk use of AI available for public inspection. HRI researchers should make these types of information available both so that developers building off their work can easily accommodate legally mandated reporting guidelines,

and so that lawyers, policymakers, and regulators can effectively monitor the advances being made by the research community and the legally relevant effects those advances may have.

Although the research community will necessarily iterate on the best mechanisms for providing transparency, the suggestions in the preceding section are intended to establish a baseline level of information appropriate for the stakeholders discussed here. To make the discussion more concrete, we next present a brief example of a fictional research project and associated reporting.

5 Reporting Example for Researchers

In this example, we consider a research project in which a robot/person dyad are performing a manipulation task together. The robot is providing additional manipulation capabilities, serving as a "third hand." The user provides spoken instructions describing what the robot should be doing, which may be simple (e.g., "move the part down a couple of inches"), or arbitrarily complex (e.g., "finish assembling this while I hold it"), with no predefined constraints on the language's level of abstraction. The system performs language understanding and interpretation by sending commercial LLM a sequence of images of the interaction and the human's spoken instructions, and prompting for what action(s) the robot should take (so it may return, for example, "move gripper 20cm left"). When reporting on this work, the researcher should answer the following questions:

(1) What model are you using, and what version of that model?
(2) Why are you using an LLM, and why that particular model?
(3) How is the LLM component evaluated? Is there a separate subsystem evaluation of performance, or is it part of the evaluation of the overall system?
(4) Are you using the LLM as a Scarecrow (i.e., as a "brainless" module in which the LLM provides a stopgap solution rather than a theoretically principled, empirically justified, and safe solution), or are you advocating for its use as part of a deployable solution?
(5) What role is the LLM serving in the architecture, and how is it integrated into that architecture?
(6) What are the ethical implications stemming from your choice to use an LLM, and from the specific way in which you are using that LLM? In particular, those using LLMs should clearly acknowledge the environmental and sociological impacts, the privacy and reliability risks, and the ways those risks specifically manifest and are accentuated or ameliorated within the research context being explored.

The inset box shows how the answers to these questions might be included in a publication.

> The system described in this work depends on the use of a Large Vision and Language Model for interpreting human instructions and selecting robot actions based on those inputs. The VLM was a separate API that was queried to find a mapping from human instructions to possible actions. For the implemented system, we used OpenAI's GPT, specifically model gpt-4o-2024-08-06. No fine-tuning was performed, and no seeds were specified. This model was selected based on its accessibility and performance on a range of queries obtained during a Wizard of Oz pilot study. We performed all queries during December 11–15, 2024. We prompt the model with the following (example instruction included):
>
> ```
> PROMPT: You are an interpreter of human instructions for basic tasks. You are
> working with a human to jointly perform a simple collaborative task. In this task
> you are a robot working with a human to build a slot together model. For a given
> statement determine if the statement is directed to the robot, is not a request or is
> not an action. If it is a request or action directed to the robot, return the action
> the robot should take as a python dictionary. The dictionary has one key: "action":
> A list of the actions that the robot should take, taken from the actions list. For a
> ```

```
given statement determine what actions the robot should take. Return only a single
object from the list of objects provided. Resume using the following instruction and
the objects in the provided image.
  "instruction": 'Okay, now hand me the red part.'
  "objects" = [FrontPiece, BackPiece, LegsFront, LegsBack, Head]
  "actions" = [MOVERIGHT, MOVELEFT, MOVEUP, MOVEDOWN, MOVEFORWARD, MOVEBACKWARD,
TILTUP, TILTDOWN, ROTATELEFT, ROTATERIGHT, PICKUP, OPENHAND, CLOSEHAND, OTHER]
```

Because the information being sent to the model includes both the human instructions and an image of the workspace, we arranged the camera to show only participants' hands and arms. There are still privacy implications in the transmission of (transcribed) human speech. However, since the task involves only language about joining pieces of a toy, there is no reason for participants to provide sensitive information, and experimenters were prepared to intervene in any case in which they did so. The overall system evaluation included questions such as "I felt like the robot understood what I wanted," but did not include a separate assessment of the VLM's performance.

From a social justice perspective, we identify two possible sources of risk. First, the model is receiving images that include participants' hands and arms, and skin tone may affect model behavior [8, 58, 59]. Participant instructions may also convey demographic information in a way that affects the system's performance [24, 50]. To minimize these risks, the prompts were designed to be task-oriented and focus on the objects in the workspace. GPT was prompted to focus on the human instructions and not the instructors. Transcribing speech to text may also reduce the impact of acoustic features such as accents.

Environmentally, although detailed figures are difficult to obtain, a conservative estimate is that each of our GPT inferences is responsible for approximately 0.047 kilowatt-hours of electricity [68]. As our study had 20 participants who issued an average of 16.3 instructions each, we estimate that the electricity usage of our study was roughly 15.3 kWh. Our proposed deployment model is to evaluate the type of instructions participants tend to use and transition to a smaller, local model tuned for those specific use cases, reducing both environmental costs and privacy concerns.

Without the use of a vision and language model, this work—which does not focus primarily on the development of an NLP interface—would have relied on simpler linguistic interactions, possibly with the use of a fixed lexicon of commands, reducing the participants' freedom to convey a variety of requirements to the robot helper. It is intuitive to expect that users would have ended up using simpler commands in this case; however, ablation studies were not performed.

Sample report on LLM usage in a hypothetical research project.

6 Next Steps

In this paper, we explored the "why?" and the "how?" of the ways researchers depend on and can provide transparency about the uses of LLMs in robot architectures for five key user groups to whom researchers owe an informative reporting responsibility: other researchers, engineers, decision makers, lawyers/regulators, and end users. However, there are a number of related considerations that must also be discussed to make clear the boundaries of our analysis.

6.1 Transparency for Additional Stakeholders

The need to be transparent does not only apply to the HRI research community. Frequently, the ultimate goal of HRI research is to understand how to build robots that can be deployed in a variety of situations in which they might supply assistance, provide work, or otherwise contribute to meeting the needs of people. Successful deployments require pipelines of programmers, engineers, manufacturers, marketers, and support personnel, among others. It is unrealistic to expect that every person involved in such an endeavor will be fully informed about the technical underpinnings

of a developed system. However, the goal is that every person involved will understand the shape of the **technological solution** and the **risks that may be involved**. As such, it becomes necessary to identify what information should be provided to whom, and what information may be expected to be passed through to other stakeholders to meet their own information needs. As an example, sales personnel need to understand enough about how LLM-based systems work and what the associated risks are that they can convey this information to potential customers.

Even before deployment, some of the same requirements apply to smaller-scale commercial efforts; for example, a startup trying to develop a robot system for hospital use should be aware of, sensitive to, and transparent about questions of privacy and information flow. At the opposite end of the spectrum, large companies that are responsible for developing, training, and publicizing very large models should be aware of the potential uses discussed in this article, and should make every effort to be as transparent as possible for the sake of avoiding problems such as test data contamination [4, 60].

6.2 Beyond Transparency

Despite the importance of transparency, it is equally critical to note the limitations of (and ethical inculpability provided by) simply providing information about a system. In practice, technical information may be difficult to understand, while excessive information may be impossible to absorb usefully [90]. At best, poorly presented information may not allow users to provide informed consent to the use of LLMs [28]; at worst, such documentation may be used as a form of "transparency-washing," in which technology providers use tools such as end-user license agreements primarily to exculpate themselves of responsibility for user understanding [31]. Designers of robotic systems (LLM-enabled or not) **must take advantage of best practices in current digital consent research** to ensure that key information is communicated effectively. The communication tools described in this article are intended as a starting point towards this practice.

6.3 The Need for Robotics and AI Literacy Efforts

This need to move beyond transparency highlights the information needs of stakeholders like end users, decision makers, and law and policy experts. Not only do these stakeholders need "reporting" from roboticists before they make decisions as to when and how to regulate, purchase, and use LLM-enabled robots, but they moreover need a solid base of knowledge about the nature of these robots, and the broad space of opportunities and risks they present. This broad base of knowledge might thus be best enabled not only through "reporting" as traditionally construed, but also through *population-level education and tech literacy efforts.*

Efforts like AI4K12 [108] (see also [94, 114]) have emphasized the importance of integrating **awareness of ethical risks** into K-12 AI education; Duke's Cultural Competence in Computing (3C) program trains educators in how to develop coursework related to issues of identity and computing [25]; Ko and colleagues' textbook on Critically Conscious Computing seeks to center societal and ethical issues in the course of secondary computing education [51]; and the AAAS's Project 2061 [72] has presented guidelines for technology literacy efforts, which have been taken up by HRI researchers like Mott and Williams [77]. Our analysis in this paper presents not only a set of guidelines that HRI researchers should follow, but also a call to action to work with educators and education researchers to develop and disseminate these types of curricula.

An interesting extension of the concept of general education that came from our workshop was the idea of **licensure** for LLMs and robotics, in which practitioners (researchers and/or engineers) would receive formal training and licensing in these systems. Such licensing might take two forms: (1) voluntary certifications, which could serve to demonstrate expertise in the uses and risks of LLM-backed robots; and (2) professional licensing, in which such expertise would be considered a

professional expectation. Licensing has the advantages of incorporating the legal system to hold practitioners accountable and provide more transparency, and could build trust in an organization dedicated to researching and using such systems. However, it also has the challenges of requiring both time and effort, making it disproportionately harder for small businesses and individuals to get involved or gain accreditation.

6.4 Expanding Reporting Guidelines for Diverse Robot Types and Human Populations

In this paper, we proposed foundational reporting guidelines for the **general use of LLMs in HRI research**. While the guidelines presented are broadly applicable, we acknowledge that HRI encompasses a wide range of robotic applications, each with unique ethical, safety, and operational considerations. As such, this paper serves as a starting point rather than a comprehensive framework addressing every potential use case.

For example, different robot types (e.g., anthropomorphic vs. non-anthropomorphic robots) will likely require tailored approaches to reporting and ethical considerations. Moreover, the diversity of human populations interacting with robots further highlights the need for specialized reporting frameworks; for example, children, older adults, and individuals with special needs may have distinct vulnerabilities that necessitate careful consideration.

We propose that future research expand upon the foundational guidelines outlined in this paper to explore these specific contexts in greater detail. Investigating how reporting requirements should vary based on different robot types (e.g., humanoids vs. non-humanoids) and different human populations (e.g., children, older adults, or people with special needs) is a critical next step. By doing so, the HRI community can ensure that the ethical, safety, and operational needs of diverse applications are adequately addressed.

7 Conclusion & Reporting Guidelines

As seen across this paper, there are a wide array of stakeholders who deserve targeted communication about the use of LLMs in robot architectures. Each of these stakeholder groups has their own motivations for such reporting, their own reporting needs, and their own ideal forms of reporting. However, we wish to end this paper by providing, as a primary takeaway, a succinct summary of the most important information that HRI researchers using LLMs should report to other researchers in the context of their scientific papers.

As a first step towards meeting the broad range of reporting guidelines described across this paper, we recommend that HRI researchers, at minimum, report, in research papers that leverage LLMs as part of larger robot architectures, the answers to the six questions listed in Section 5, which we repeat here for clarity:

(1) What model are you using, and what version of that model?
(2) Why are you using an LLM, and why that particular model?
(3) How is the LLM component evaluated? Is there a separate subsystem evaluation of performance, or is it part of the evaluation of the overall system?
(4) Are you using the LLM as a Scarecrow (i.e., as a "brainless" module in which the LLM provides a stopgap solution rather than a theoretically principled, empirically justified, and safe solution), or are you advocating for its use as part of a deployable solution?
(5) What role is the LLM serving in the architecture, and how is it integrated into that architecture?
(6) What are the ethical implications stemming from your choice to use an LLM, and from the specific way in which you are using that LLM? In particular, those using LLMs should clearly acknowledge the environmental and sociological impacts, the privacy and reliability risks,

and the ways those risks specifically manifest and are accentuated or ameliorated within the research context being explored.

By making this minimal set of information available to other HRI researchers, we can start creating a culture of appropriate and mutual transparency toward the range of stakeholders to whom the field of HRI owes a reporting burden.

Acknowledgments

We would like to thank the attendees of the HRI 2024 Workshop on *Scarecrows in Oz: Large Language Models in Human-Robot Interaction* for their participation in informing discussions on these topics. Tom Williams' contributions were funded in part by NSF CAREER Award IIS-2044865 and in part by Office of Naval Research award N00014-21-1-2418. Ruchen Wen and Cynthia Matuszek's work was supported in part by NSF grants IIS-2024878 and IIS-2145642, and this material is also based on research that is in part supported by the Army Research Laboratory, Grant No. W911NF2120076. Eike Schneiders' contributions was supported by the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1] UKRI Trustworthy Autonomous Systems Hub and Responsible AI UK [grant number EP/Y009800/1]. Casey Kennington's contributions were supported in part by NSF CAREER award IIS/EPSCoR-2140642.

References

[1] Philipp Allgeuer, Hassan Ali, and Stefan Wermter. 2024. When robots get chatty: Grounding multimodal human-robot conversation and collaboration. In *International Conference on Artificial Neural Networks*. Springer, 306–321.

[2] Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, et al. 2021. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436.

[3] Stefan Baack, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, Maximilian Gahntz, Paul Keller, Pierre-Carl Langlais, Greg Lindahl, Sebastian Majstorovic, Nik Marda, Guilherme Penedo, Maarten Van Segbroeck, Jennifer Wang, Leandro von Werra, Mitchell Baker, Julie Belião, Kasia Chmielinski, Marzieh Fadaee, Lisa Gutermuth, Hynek Kydlíček, Greg Leppert, EM Lewis-Jong, Solana Larsen, Shayne Longpre, Angela Oduor Lungati, Cullen Miller, Victor Miller, Max Ryabinin, Kathleen Siminyu, Andrew Strait, Mark Surman, Anna Tumadóttir, Maurice Weber, Rebecca Weiss, Lee White, and Thomas Wolf. 2025. Towards Best Practices for Open Datasets for LLM Training. arXiv:2501.08365 [cs.CY] https://arxiv.org/abs/2501.08365

[4] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta.

[5] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications* (2024), 100545.

[6] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security* 6, 1 (2023), 1–52.

[7] 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[8] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1244.

[9] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy?. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2280–2292.

[10] Susan A Brown, Anne P Massey, Mitzi M Montoya-Weiss, and James R Burkman. 2002. Do I really have to? User acceptance of mandated technology. *European journal of information systems* 11, 4 (2002), 283–295.

[11] 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 (2020), 1877–1901.

[12] Wasja Brunotte, Larissa Chazette, Lukas Kohler, Jil Klunder, and Kurt Schneider. 2022. What about my privacy? helping users understand online privacy policies. In *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*. 56–65.

[13] Wasja Brunotte, Alexander Specht, Larissa Chazette, and Kurt Schneider. 2023. Privacy explanations–A means to end-user trust. *Journal of Systems and Software* 195 (2023), 111545.

[14] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 5253–5270.

[15] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

[16] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*. 267–284.

[17] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[18] Stephen Cave and Kanta Dihal. 2020. The Whiteness of AI. *Philosophy & Technology* 33, 4 (2020), 685–703.

[19] Guanqi Chen, Lei Yang, Ruixing Jia, Zhe Hu, Yizhou Chen, Wei Zhang, Wenping Wang, and Jia Pan. 2024. Language-Augmented Symbolic Planner for Open-World Task Planning. In *Proceedings of Robotics: Science and Systems (RSS)*.

[20] Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. Comprehensive assessment of jailbreak attacks against LLMs. *arXiv preprint arXiv:2402.05668* (2024).

[21] Valeria Cirillo, Matteo Rinaldini, Jacopo Staccioli, and Maria Enrica Virgillito. 2021. Technology vs. workers: the case of Italy's Industry 4.0 factories. *Structural Change and Economic Dynamics* 56 (2021), 166–183. https://doi.org/10.1016/j.strueco.2020.09.007

[22] Mark Coeckelbergh. 2025. LLMs, Truth, and Democracy: An Overview of Risks. *Science and Engineering Ethics* 31, 1 (2025), 1–13.

[23] Colorado General Assembly. 2024. Consumer Protections for Artificial Intelligence, Bill SB24-205. https://leg.colorado.gov/bills/sb24-205, See also https://fpf.org/wp-content/uploads/2024/05/FPF-FINAL-CO-SB-205-Two-Pager-.pdf.

[24] Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American Language Bias in Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore.

[25] Jasmine DeHart. 2022. Building cultural competency for computing: identity computing lab, Duke University. *XRDS: Crossroads, The ACM Magazine for Students* 28, 4 (2022), 38–39.

[26] Yan Ding, Xiaohan Zhang, Saeid Amiri, Nieqing Cao, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. 2023. Integrating action knowledge and LLMs for task planning and situation handling in open worlds. *Autonomous Robots* 47, 8 (2023), 981–997.

[27] Wenchao Dong, Assem Zhunis, Hyojin Chin, Jiyoung Han, and Meeyoung Cha. 2024. I Am Not Them: Fluid Identities and Persistent Out-group Bias in Large Language Models. *arXiv preprint arXiv:2402.10436* (2024).

[28] Elizabeth Edenberg and Meg Leta Jones. 2019. Analyzing the legal roots and moral core of digital consent. *New Media & Society* 21, 8 (2019), 1804–1823.

[29] Jide S Edu, Jose M Such, and Guillermo Suarez-Tangil. 2020. Smart home personal assistants: a security and privacy review. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–36.

[30] Georgios Feretzakis, Evangelia Vagena, Konstantinos Kalodanis, Paraskevi Peristera, Dimitris Kalles, and Athanasios Anastasiou. 2025. GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet* 17, 4 (2025), 151.

[31] Catherine Flick. 2009. *Informed consent in information technology: Improving user experience*. Ph. D. Dissertation. Charles Stuart University.

[32] Pedro Rubim Borges Fortes, Pablo Marcello Baquero, and David Restrepo Amariles. 2022. Artificial intelligence risks and algorithmic regulation. *European Journal of Risk Regulation* 13, 3 (2022), 357–372.

[33] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[34] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[35] Nina Gerber, Paul Gerber, and Melanie Volkamer. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & security* 77 (2018), 226–261.
[36] Sara Gerke. 2023. "Nutrition Facts Labels" for Artificial Intelligence/Machine Learning-Based Medical Devices-The Urgent Need for Labeling Standards. *Geo. Wash. L. Rev.* 91 (2023), 79.
[37] Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2025. AEGIS2. 0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails. *arXiv preprint arXiv:2501.09004* (2025).
[38] Bengio Yoshua Goodfellow, Ian and Aaron Courville. 2016. *Deep Learning*. MIT press.
[39] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems* 24, 2 (2009), 8–12.
[40] Cari Beth Head, Paul Jasper, Matthew McConnachie, Linda Raftree, and Grace Higdon. 2023. Large language model applications for evaluation: Opportunities and ethical implications. *New directions for evaluation* 2023, 178-179 (2023), 33–46.
[41] Areum Hong, Changi Nam, and Seongcheol Kim. 2020. What will be the possible barriers to consumers' adoption of smart home services? *Telecommunications Policy* 44, 2 (2020), 101867.
[42] Silvia Izquierdo-Badiola, Gerard Canal, Carlos Rizzo, and Guillem Alenyà. 2024. PlanCollabNL: leveraging Large Language Models for adaptive plan generation in human-robot collaboration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 17344–17350.
[43] Yousra Javed, Shashank Sethi, and Akshay Jadoun. 2019. Alexa's voice recording behavior: A survey of user understanding and awareness. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*. 1–10.
[44] Wenxin Jiang, Nicholas Synovic, Matt Hyatt, Taylor R Schorlemmer, Rohan Sethi, Yung-Hsiang Lu, George K Thiruvathukal, and James C Davis. 2023. An empirical study of pre-trained model reuse in the hugging face deep learning model registry. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2463–2475.
[45] Dzieza Josh. 2023. Inside the AI Factory. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots. Accessed: 2024-06-22.
[46] Tharindu Kaluarachchi, Andrew Reis, and Suranga Nanayakkara. 2021. A review of recent deep learning approaches in human-centered machine learning. *Sensors* 21, 7 (2021), 2514.
[47] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' mental models of AI: An item response theory approach. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (FAccT*. 1723–1734.
[48] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 371–380.
[49] Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 371–380.
[50] Shana Kleiner, Jessica A Grieser, Shug Miller, James Shepard, Javier Garcia-Perez, Nick Deas, Desmond U Patton, Elsbeth Turcan, and Kathleen McKeown. 2024. Unmasking camouflage: exploring the challenges of large language models in deciphering African American language & online performativity. *AI and Ethics* (2024), 1–9.
[51] Amy J Ko, Anne Beitlers, Brett Wortzman, Matt Davidson, Alannah Oleson, Mara Kirdani-Ryan, Stefania Druga, and Jayne Everson. 2022. Critically conscious computing: methods for secondary education. *Online. https://criticallyconsciouscomputing. org* (2022).
[52] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.
[53] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security* 64 (2017), 122–134.
[54] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, et al. 2024. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science* 11, 5 (2024), 230859.
[55] Younes Lakhnati, Max Pascher, and Jens Gerken. 2024. Exploring a GPT-based large language model for variable autonomy in a VR-based human-robot teaming simulation. *Frontiers in Robotics and AI* 11 (2024). https://doi.org/10.3389/frobt.2024.1347538
[56] Donghyeok Lee, Christina Todorova, and Alireza Dehghani. 2024. Ethical Risks and Future Direction in Building Trust for Large Language Models Application under the EU AI Act. In *Proceedings of the 2024 Conference on Human*

Centred Artificial Intelligence—Education and Practice. 41–46.

[57] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[58] Messi H.J. Lee and Soyeon Jeon. 2024. Vision-Language Models Represent Darker-Skinned Black Individuals as More Homogeneous than Lighter-Skinned Black Individuals. *arXiv preprint arXiv:2412.09668* (2024).

[59] Messi H.J. Lee, Jacob M Montgomery, and Calvin K Lai. 2024. More Distinctively Black and Feminine Faces Lead to Increased Stereotyping in Vision-Language Models. *arXiv preprint arXiv:2407.06194* (2024).

[60] Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18471–18480.

[61] Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. 2023. Making AI less "thirsty"': Uncovering and addressing the secret water footprint of AI models. *arXiv preprint arXiv:2304.03271* (2023).

[62] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems* 35 (2022), 31199–31212.

[63] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-centered privacy research in the age of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

[64] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning. In *International conference on machine learning*. PMLR, 19565–19594.

[65] Zihao Li. 2023. Why the European AI Act transparency obligation is insufficient. *Nature Machine Intelligence* 5, 6 (2023), 559–560.

[66] Baisong Liu, Daniel Tetteroo, and Panos Markopoulos. 2022. A systematic review of experimental work on persuasive social robots. *International Journal of Social Robotics* 14, 6 (2022), 1339–1378.

[67] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. 2023. Constrained decision transformer for offline safe reinforcement learning. In *International Conference on Machine Learning*. PMLR, 21611–21630.

[68] Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment?. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 85–99.

[69] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 482–491.

[70] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. LLM Dataset Inference: Did you train on my dataset?. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 124069–124092. https://proceedings.neurips.cc/paper_files/paper/2024/file/e01519b47118e2f51aa643151350c905-Paper-Conference.pdf

[71] Nathan Malkin, David Wagner, and Serge Egelman. 2022. Can Humans Detect Malicious Always-Listening Assistants? A Framework for Crowdsourcing Test Drives. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[72] William F McComas. 2014. Benchmarks for science literacy. *The Language of Science Education: An Expanded Glossary of Key Terms and Concepts in Science Teaching and Learning* (2014), 12–12.

[73] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. 121–135.

[74] Charles W Mills. 2005. Kant's Untermenschen. In *Race and Racism in Modern Philosophy*. Cornell University Press, 169–193.

[75] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI* 10 (2023), 1271610.

[76] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[77] Terran Mott and Tom Williams. 2023. Community Futures With Morally Capable Robotic Technology. In *Workshop on Perspectives on Moral Agency in Human-Robot Interaction at HRI*.

[78] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. 2022. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing* 73 (2022), 102231.

[79] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. In *The Twelfth International Conference on Learning Representations*.

[80] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[81] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–21.

[82] Action Network. 2024. Commit to Fair Work: 5 Simple Steps For a Just Platform Economy. https://actionnetwork.org/forms/commit-to-fair-work-5-simple-steps-for-a-just-platform-economy. Accessed: 2024-06-22.

[83] Terrence Neumann, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. 2024. Diverse, but Divisive: LLMs Can Exaggerate Gender Differences in Opinion Related to Harms of Misinformation. *arXiv preprint arXiv:2401.16558* (2024).

[84] Linus Nwankwo and Elmar Rueckert. 2024. The Conversation is the Command: Interacting with Real-World Autonomous Robots Through Natural Language. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 808–812.

[85] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating Affective Use and Emotional Well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).

[86] Jeremias Prassl. 2018. *Humans as a service: The promise and perils of work in the gig economy*. Oxford University Press.

[87] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access* 12 (2024), 26839–26874.

[88] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[89] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 23–72.

[90] Neil Richards and Woodrow Hartzog. 2018. The pathologies of digital consent. *Wash. UL Rev.* 96 (2018), 1461.

[91] Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[92] Kantwon Rogers, Reiden John Allen Webber, Geronimo Gorostiaga Zubizarreta, Arthur Melo Cruz, Shengkang Chen, Ronald C Arkin, Jason Borenstein, and Alan R Wagner. 2024. What Should a Robot Do? Comparing Human and Large Language Model Recommendations for Robot Deception. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 906–910.

[93] Sandra C Sandoval, Christabel Acquaye, Kwesi Cobbina, Mohammad Nayeem Teli, and Hal Daumé III. 2025. My LLM might Mimic AAE–But When Should it? *arXiv preprint arXiv:2502.04564* (2025).

[94] Ismaila Temitayo Sanusi and Sunday Adewale Olaleye. 2022. An insight into cultural competence and ethics in K-12 artificial intelligence education. In *2022 IEEE global engineering education conference (EDUCON)*. IEEE, 790–794.

[95] Ranjan Sapkota, Shaina Raza, and Manoj Karkee. 2025. Comprehensive analysis of transparency and accessibility of chatgpt, deepseek, and other sota large language models. *arXiv preprint arXiv:2502.18505* (2025).

[96] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (2021), 20539517211053712.

[97] Eva-Maria Schomakers, Hannah Biermann, and Martina Ziefle. 2021. Users' Preferences for Smart Home Automation – Investigating Aspects of Privacy and Trust. *Telematics and Informatics* 64 (2021), 101689. https://doi.org/10.1016/j.tele.2021.101689

[98] Francesco Semeraro, Alexander Griffiths, and Angelo Cangelosi. 2023. Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing* 79 (2023), 102432.

[99] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11523–11530.

[100] Mary Anne Smart. 2021. Addressing privacy threats from machine learning. *arXiv preprint arXiv:2111.04439* (2021).

[101] Nathalie A Smuha. 2021. From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology* 13, 1 (2021), 57–84.
[102] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. LLM-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2998–3009.
[103] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).
[104] Jovan Stojkovic, Esha Choukse, Chaojie Zhang, Inigo Goiri, and Josep Torrellas. 2024. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. *arXiv preprint arXiv:2403.20306* (2024).
[105] Lingfeng Sun, Devesh K Jha, Chiori Hori, Siddarth Jain, Radu Corcodel, Xinghao Zhu, Masayoshi Tomizuka, and Diego Romeres. 2024. Interactive planning using large language models for partially observable robotic tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 14054–14061.
[106] Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. 2024. Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science. *arXiv preprint arXiv:2402.04247* (2024).
[107] Kassym-Jomart Tokayev. 2023. Ethical implications of large language models: a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics* 8, 9 (2023), 17–33.
[108] David Touretzky, Fred Martin, Deborah Seehorn, Cynthia Breazeal, and Tess Posner. 2019. Special session: AI for K-12 guidelines initiative. In *Proceedings of the 50th ACM technical symposium on computer science education*. 492–493.
[109] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171* (2024).
[110] Twilio. 2023. AI Nutrition Label. https://nutrition-facts.ai/.
[111] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion?. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 36–45.
[112] Alan R Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the robotic age. *Commun. ACM* 61, 9 (2018), 22–24.
[113] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802* (2024).
[114] Randi Williams. 2021. How to train your robot: Project-based ai and ethics education for middle school classrooms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1382–1382.
[115] Tom Williams. 2024. Understanding Roboticists' Power through Matrix Guided Power Analysis. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 46–56.
[116] Tom Williams. 2025 **(Forthcoming)**. *Degrees of Freedom: On Robotics and Social Justice*. MIT Press.
[117] Tom Williams, Cynthia Matuszek, Kristiina Jokinen, Raj Korpan, James Pustejovsky, and Brian Scassellati. 2023. Voice in the machine: Ethical considerations for language-capable robots. *Commun. ACM* 66, 8 (2023), 20–23.
[118] Tom Williams, Cynthia Matuszek, Ross Mead, and Nick Depalma. 2024. Scarecrows in Oz: The Use of Large Language Models in HRI. *ACM Transactions on Human-Robot Interaction* 13, 1 (2024), 1–11.
[119] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The Confucian Matador: Three Defenses against the Mechanical Bull. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. 25–33.
[120] Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. 2024. A New Era in LLM Security: Exploring Security Concerns in Real-World LLM-based Systems. *arXiv preprint arXiv:2402.18649* (2024).
[121] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* 47, 8 (2023), 1087–1102.
[122] Xiyang Wu, Ruiqi Xian, Tianrui Guan, Jing Liang, Souradip Chakraborty, Fuxiao Liu, Brian Sadler, Dinesh Manocha, and Amrit Singh Bedi. 2024. On the Safety Concerns of Deploying LLMs/VLMs in Robotics: Highlighting the Risks and Vulnerabilities. *arXiv preprint arXiv:2402.10340* (2024).
[123] Zihao Wu, Peng Shu, Yiwei Li, Quanzheng Li, Tianming Liu, and Xiang Li. 2024. Robot Control via Natural Instructions Empowered by Large Language Model. In *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*. Springer, 437–457.
[124] Sarah Wyer and Sue Black. 2025. Algorithmic bias: sexualized violence against women in GPT-3 models. *AI and Ethics* (2025), 1–18.
[125] Ala Yankouskaya, Areej B Babiker, Syeda WF Rizvi, Sameha Alshakhsi, Magnus Liebherr, and Raian Ali. 2025. LLM-D12: A Dual-Dimensional Scale of Instrumental and Relational Dependencies on Large Language Models. *arXiv*

preprint arXiv:2506.06874 (2025).

[126] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.

[127] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced LLM bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*. 292–299.

[128] Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. 2024. Object Recognition as Next Token Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16645–16656.

[129] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[130] Bowen Zhang and Harold Soh. 2023. Large language models as zero-shot human models for human-robot interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7961–7968.

[131] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics* (2023), 100131.

[132] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[133] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027* (2024).

[134] Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. 2023. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028* (2023).

[135] Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. Quantifying and Analyzing Entity-Level Memorization in Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19741–19749.