arXiv:2306.01941v1 [cs.HC] 2 Jun 2023

AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap

Q. Vera Liao and Jennifer Wortman Vaughan*
Microsoft Research

June 6, 2023

Abstract

The rise of powerful large language models (LLMs) brings about tremendous opportunities for innovation but also looming risks for individuals and society at large. We have reached a pivotal moment for ensuring that LLMs and LLM-infused applications are developed and deployed responsibly. However, a central pillar of responsible AI—transparency—is largely missing from the current discourse around LLMs. It is paramount to pursue new approaches to provide transparency for LLMs, and years of research at the intersection of AI and human-computer interaction (HCI) highlight that we must do so with a human-centered perspective: Transparency is fundamentally about supporting appropriate human understanding, and this understanding is sought by different stakeholders with different goals in different contexts. In this new era of LLMs, we must develop and design approaches to transparency by considering the needs of stakeholders in the emerging LLM ecosystem, the novel types of LLM-infused applications being built, and the new usage patterns and challenges around LLMs, all while building on lessons learned about how people process, interact with, and make use of information. We reflect on the unique challenges that arise in providing transparency for LLMs, along with lessons learned from HCI and responsible AI research that has taken a human-centered perspective on AI transparency. We then lay out four common approaches that the community has taken to achieve transparency—model reporting, publishing evaluation results, providing explanations, and communicating uncertainty—and call out open questions around how these approaches may or may not be applied to LLMs. We hope this provides a starting point for discussion and a useful roadmap for future research.

1 Introduction

Hugely powerful large language models (LLMs) like GPT-4, LaMDA, and LLaMA are now being deployed in applications from search engines to code generation tools to productivity suites. These generative models are widely expected to have impact across industries, changing the way we engage in tasks like writing, programming, and design, and reshaping occupations in medicine, law, marketing, education, and beyond [3, 22, 39, 40, 50, 111]. As the chair of the Federal Trade Commission put it in a recent op ed, "the full extent of generative AI's potential is still up for debate, but there's little doubt it will be highly disruptive" [93].

While the capabilities of LLMs are impressive, they also raise new risks [16, 104, 146, 188]. Language models are found to encode biases [2, 152], which risks propagating harmful discrimination, stereotypes, and exclusion at scale. They are widely known to "hallucinate" information [16, 85, 122, 135], producing outputs that are plausible—even convincing—but incorrect. They may project confidence about these hallucinated outputs, potentially contributing to automation bias, overreliance, or automation-induced complacency [148, 189]. LLMs can generate harmful, sometimes toxic content, including hate speech and offensive language [16, 187], or reveal sensitive information that threatens privacy or security. They can contribute—both intentionally and unintentionally—to the spread of misinformation [28, 101, 201]. And in the longer term, LLMs may lead to environmental harms [16] as well as socioeconomic harms, including the displacement and deskilling of workers across industries [187].

Given the anticipated impact that LLMs will have on both our day-to-day lives and society at large, it is critical that LLMs and LLM-infused applications be developed and deployed responsibly. One

---

[*] equal contribution

central component of responsible AI development and deployment is *transparency*: enabling relevant stakeholders to form an appropriate understanding of a model or system's capabilities, limitations, how it works, and how to use or control its outputs. Developers of LLMs cannot debug their models, responsibly assess whether they are ready to launch, and enforce responsible and safe usage policies for their models without some understanding of their behavior and performance on different tasks. Business decision-makers, designers, and developers building LLM-infused applications must be able to understand the LLM's capabilities and limitations in order to ideate and make decisions about whether, where, and how to use the model—potentially including how to fine-tune, prompt, or otherwise adapt the model to better fit their use case. End-users must be able to form a sufficiently accurate understanding of LLM-infused applications to control the application's behavior and achieve appropriate levels of trust and reliance. People impacted by LLMs or LLM-infused applications may require transparency in order to understand their options for recourse. Additionally, given the speed at which powerful new LLMs and their applications are being released and the growing concerns over potential harms, we should expect to see an increased demand for transparency around their development and inner workings from policymakers and third-party auditors aiming to regulate and oversee their use.

In recent years, we have witnessed the creation of a whole research field at the intersection of AI and human-computer interaction (HCI) that is focused on developing and evaluating different approaches to achieve transparency. These approaches range from frameworks for documenting models and the datasets they are trained on [e.g., 6, 15, 37, 56, 74, 139] to techniques for producing explanations of individual model outputs [e.g., 100, 128, 155, 157, 176] to approaches for communicating uncertainty [e.g., 19, 42, 183] and beyond. There is no one-size-fits-all solution. In the case of LLMs, the needs of an application developer engaging in ideation are probably different from those of a writer who is using an LLM-infused application to edit a novel or a public figure who is concerned with the way in which an LLM-infused search engine is presenting misinformation about their life. In our own work [116, 180], we have argued for the importance of taking a human-centered perspective on transparency—designing and evaluating transparency approaches with stakeholders and their goals in mind. We believe that this is even more important in the era of LLMs, when the diversity of stakeholders and their experience levels, contexts, goals, and transparency needs, is greater than ever.

In this paper, we map out a human-centered research roadmap for transparency in this new era. We first reflect on the unique challenges that arise in providing transparency for LLMs compared with smaller-scale, more specialized models that have traditionally been the focus of AI transparency research. We reflect on lessons learned from HCI and Responsible AI/FATE (fairness, accountability, transparency, and ethics) research that centers on human needs of, interactions with, and impact from AI transparency. We then lay out common approaches, including techniques and artifacts, that the community has taken to achieve transparency and call out open questions around how these might be applied to LLMs.

We note that there is no agreed-upon definition of transparency, and indeed, transparency has been recognized as a multi-faceted concept. In this paper, we adopt a focus on *informational* transparency—essentially, what information about a model (or system building on that model) should be disclosed to enable appropriate understanding—which has been emphasized within the machine learning (ML) research community and in industry practice, though we note that there are other perspectives, such as the normative, relational, and social dimensions of transparency, that have been studied in the broader literature [51, 136]. Some of the approaches we cover, such as model reporting, are primarily aimed at supporting a *functional* understanding of *what* the model (or system) can do, often by exposing the goals, functions, overall capabilities, and limitations. Others, like the explanations frequently explored in the explainable AI (XAI) and interpretable ML communities, are primarily aimed at supporting a *mechanistic* understanding of *how* the model (or system) works, by disclosing the parts and processes [126]. We believe that both understandings play important roles and the appropriate form of transparency in any given context will depend on the stakeholder and the goal that they wish to achieve.

Finally, we note that many of the challenges, lessons learned, potential approaches, and open problems that we call out in this paper apply not only to LLMs but to other large-scale generative models, including multimodal models that allow for both textual and visual input or output. While we adopt the narrower focus on LLMs for simplicity, we encourage additional research on transparency for these other models.

2 What Makes Transparency for LLMs Challenging?

To ground the discussion in the remainder of the paper, we first explore the unique characteristics of LLMs and the emerging patterns of their usage that are likely to make it more challenging to achieve transparency compared with the smaller-scale, specialized models that AI transparency research has traditionally dealt with. We start by providing some brief background on LLMs and establishing some terminology that we will use in the rest of the paper.

2.1 Background on LLMs

An LLM, like any language model, predicts the conditional probability of a token—which might be a character, word, or other string—given its preceding context and, in the case of bidirectional models, its surrounding context [17, 151]. While earlier language models were based on simple ideas like n-grams, present-day LLMs are based on modern neural network self-attention architectures like the transformer [179] with hundreds of billions or even more than a trillion parameters [54]. While earlier models were trained on datasets of moderate size, LLMs are trained on datasets of massive scale, with hundreds of billions or even more than a trillion tokens [23, 73], requiring many orders of magnitude more compute time. This makes LLMs vastly more sophisticated and expressive than their predecessors.

While a basic pre-trained LLM model can be viewed as a "general-purpose" next-word predictor, LLMs can be adapted or customized to exhibit or suppress specific behaviors or to perform better on specific tasks like text summarization, question answering, or code generation. One common approach is fine-tuning, in which the model's parameters are updated based on additional, specialized data [e.g., 41, 77, 109, 150]. A popular technique for fine-tuning is reinforcement learning from human feedback (RLHF) in which human preferences are used as a reward signal [35, 147]. Another approach is prompting or prompt engineering, in which natural-language prompts—often containing examples of tasks (for few-shot prompting/in-context learning) or demonstrations of reasoning (for chain-of-thought prompting)—are provided to the model to alter its behavior without making any changes to the model's internal parameters [e.g., 26, 124, 163, 186]. The resulting adapted models can then be incorporated into applications such as chatbots, search engines, or productivity tools. It is important to note that the party adapting the model and building the application is frequently not the same party who built the underlying pre-trained LLM, and may only be able to access the LLM through an API. When we talk about transparency, we must keep in mind whether we are referring to transparency about the pre-trained LLM, transparency about the adapted LLM, or transparency about the application using the pre-trained or adapted model (LLM-infused application). We aim to call out which of these we are referring to when it is not clear from the context.

2.2 Challenges for Achieving Transparency

There are several characteristics of LLMs and their usage that pose challenges for transparency. The list we lay out here is not meant to be exhaustive, but to provide context for later discussion.

New and Complex Model Capabilities and Behaviors. LLMs can perform an astonishingly wide variety of tasks in different contexts [22]. Unlike classical machine learning models where there is typically a well-defined structure of inputs and outputs, LLMs are more flexible. The capabilities of LLMs—sometimes also referred to as use cases [147] or tasks [114] in the literature—include question answering, dialogue generation, sentence completion, summarization, paraphrasing, elaboration, rewriting, classification, and more. Researchers are now additionally identifying "emergent capabilities" of LLMs—like performing arithmetic or chain-of-thought reasoning—that are not present in smaller-scale models but emerge at scale [185]. Furthermore, as described above, the precise behavior and capabilities of an LLM can be steered through approaches like fine-tuning and prompting. This adaptation can be done by the original model creators, application developers, or even end-users. All of this contributes to "capability unpredictability" [54], the idea that an LLM's capabilities cannot be fully anticipated, even by the model's creators, until its behavior on certain input is observed.

Additionally, present-day LLMs exhibit unreliable behaviors. They change unpredictably with updates. Depending on the sampling strategy used [75], outputs can be non-deterministic in the sense that the same prompt leads to a different response when input to the model again. They can misinterpret a prompt in unpredictable ways. And they may not respond consistently to a type of prompt, making the behavior of adapted models difficult to predict. These unreliable behaviors can make it challenging, if not impossible, to gain a generalized understanding of the model's behavior.

Massive and Opaque Architectures. Given the complexity and massive scale of present-day LLMs, there are currently no techniques that would provide us with a complete picture of the knowledge reflected in a model or the reasoning that is used to produce its output [24]. The mechanism of the transformer architecture underpinning LLMs is yet to be fully understood, even among experts, and some techniques that initially appear promising for interpreting the behavior of LLMs, such as looking at attention weights or perturbing inputs, can be misleading [21, 52, 82]. A more unique challenge with LLMs is the massive scale of the training data and diverse sources from which it is pulled—for example, Common Crawl and Wikipedia, often with no specific topics or formats targeted, and no thorough documentation about how the dataset was developed [94]. This makes it challenging, if not impossible, to understand what went into an LLM's training. There are currently no established answers, even in the research community, to questions such as precisely how and why these models work as well as they do, why they can or cannot perform certain tasks, and how different characteristics of the training data impact model capabilities.

Proprietary Technology. An elephant in the room that will inevitably inhibit attempts at transparency for LLMs is the proprietary nature of the models. Currently, while the efforts for developing open-source LLMs are growing [160, 174], most of the powerful LLMs were developed at large technology companies or other non-academic organizations. They are either released through APIs or completely proprietary, making it impossible to access their inner workings (e.g., weights and parameters). In many cases, details such as the size, make-up and provenance of the training data, the number of parameters, and the resources required to train the model are also not shared publicly. In essence, then, such models can only be probed in a black-box manner, which may not be sufficient to meet the transparency requirements for application developers, regulators, or other parties. Addressing this fundamental challenge may not be possible without policy and regulatory efforts that enforce transparency requirements on LLM creators and providers.

New and Complex Applications. End-users may not interact with LLMs directly, but rather through LLM-infused applications. Emerging applications include general and specialized chatbots, web search, programming assistants, productivity tools such as for writing support or presentation generation, and text analysis tools such as for customer insights discovery. As LLMs' capabilities continue to be discovered, we can only expect the number and variety of LLM-infused applications to grow. While any opacity of the model will likely trickle down to hinder the transparency of the applications built on these models, as mentioned above, the transparency requirements for LLM-infused applications will be different from the model as they serve a different set of stakeholders with different goals. Furthermore, just as the models themselves are flexible, the use cases for LLM-infused applications can be flexible and open-ended. For example, an LLM-infused search engine may be used to plan a trip, research a report, or write a poem—use cases that reflect different needs in terms of accuracy, verifiability of output, and likely the required transparency approaches.

To further complicate transparency around LLM-infused applications, such applications may not be built on a single LLM, but may involve many interacting models. For example, auxiliary LLMs can be used to review or grade the output of a primary LLM, or to expand its capabilities. LLMs can be embedded in a complex system to operate other models or external services, for example through (arbitrarily complex) plugins, allowing them to perform tasks like ordering groceries or book flights with no human in the loop. An application may also include other components like input or output filters. For example, an LLM-infused search engine may rely on results obtained from a traditional search engine to "ground" its responses. Changes to any component can change the behavior of the application, making it more difficult to understand its behavior. Approaches to transparency must therefore take into account all components and how they fit together rather than focusing on a single LLM in isolation.

Expanded and Diverse Stakeholders. As the number of LLM-infused applications grows and popular applications such as LLM-infused search engines expand their user bases, a larger number of people—diverse along many dimensions—will interact with or be impacted by LLMs. Research in AI transparency typically considers stakeholder groups like data scientists and model developers, business decision-makers, regulators and auditors, end-users, and impacted groups (i.e., the people who are directly or indirectly affected by a model or application) [76, 116, 180]. The use of LLMs may introduce new stakeholder groups with unique transparency needs. For example, it is increasingly common for product teams to have dedicated prompt engineers—a role that, until recently, did not exist—to streamline tasks, evaluate models, or contribute to model adaptation. As another example, as LLMs are increasingly used for productivity support to augment people's writing, we must consider both the creators of LLM-assisted articles and the consumers of these articles as "users" of the LLM's outputs and support both groups' transparency needs. Meanwhile, we must support any subjects referred to in the articles as "impacted groups."

As the pre-trained nature of LLMs lowers the barrier to using and building on AI capabilities, we believe application developers—including those working on model adaptation—will become a significant group and diverse in itself, potentially including developers, entrepreneurs, product managers, designers, or essentially anyone. In some cases, the line between application developers and end-users may be blurred. Consider, for example, a writer who experiments with using an LLM for writing support. This writer might benefit from model transparency to assess the LLM's suitability for different writing tasks and identify effective ways to adapt the model for each task.

Recent research has begun to inquire about the ecosystem of LLMs and the roles in it [22], from data creation, curation, model training, and model adaptation through to deployment. Identifying these LLM stakeholder roles and supporting their role-, task- and context-specific transparency needs will be of primary importance for the AI transparency research community.

Rapidly Evolving and Often Flawed Public Perception. Effective approaches to transparency should take into account the receivers' existing perception of what the model or system can or cannot do, and how it works—often referred to as their mental model [58, 87, 144]. This is especially challenging for LLMs as their public perception is still evolving and shaped by complex mechanisms including mass media, marketing campaigns, ongoing events, and design choices of popular LLM-infused applications. The natural language modality also contributes to a unique set of challenges: people may be more likely to assign human-like attributes to the model and have corresponding expectations [143], and even subtle language and communication cues can have profound impact on people's mental model [1]. Recent studies show that people already have flawed mental models about LLMs, such as incorrect perceptions of how their output differs from human-written texts [83]. Interacting with LLMs with a flawed mental model can lead to misuse, unsafe use, over- and under-reliance, deception, privacy and security threats, and other interaction-based harms [187]. Flawed public perceptions can be attributed to a lack of accurate, comprehensive, and responsible information. In addition to incorporating transparency approaches, the organizations creating LLMs and LLM-infused applications and the research community more broadly should reflect on the implications of the way they communicate with the public. For example, the use of ill-defined, catch-all phrases such as "general-purpose model" or inappropriate anthropomorphizing may hinder accurate public perception of LLMs.

Organizational Pressure to Move Fast and Deploy at Scale. Lastly, we note that there are organizational challenges that may hinder the development and adoption of transparency approaches beyond the proprietary nature of LLM. Responsible AI efforts are often in tension with pressures to release products quickly and to scale up across geographies, use cases, and user bases [130, 131, 153], a kind of "scale thinking" [70]. Given the speed at which research and product breakthroughs are occurring and the vast financial stakes, companies are incentivized to move at a pace that is unusual to witness even in the technology industry to be first to market—what some media outlets have dubbed an "AI race" or "AI arms race" [34, 63]. The organizations building LLMs and LLM-infused applications will need to take extra steps to ensure that transparency and other responsible AI considerations are not lost in the process, which may require enhanced internal governance or external regulatory requirements in addition to organizational incentives for the individuals working in this space.

3 What Lessons Can We Learn from Prior Research?

Despite the new challenges that LLMs pose, there are lessons we can learn from the HCI and Responsible AI/FATE research communities, which tend to take a human-centered perspective on transparency, centering human needs and impact. We reflect on these lessons next.

3.1 Transparency as a Means to Many Ends: A Goal-Oriented Perspective

Within the HCI community, researchers have attempted to guide the development and evaluation of transparency approaches by digging into the reasons why people seek information [107, 169]. This goal-oriented perspective resonates with studies of human explanations from the social sciences [127, 138], where it is recognized that seeking explanations and achieving understanding is often a means to an end for downstream cognitive tasks like learning, decision-making, trust development, and diagnosis.

This goal-oriented perspective has led to works developing taxonomies of common goals that people seek explanations for [117, 118, 169] and empirical studies to delineate common transparency goals of stakeholders groups such as data scientists [18, 76] and designers [119]. For example, Suresh et al. [169] lay out a set of common goals that people seek AI explanations for, including improving a model, ensuring regulatory compliance, taking actions based on model output, justifying actions influenced by the model, understanding data usage, learning about a domain, and contesting model decisions.

The goal-oriented perspective has several practical implications for developing human-centered transparency approaches. First, whether a transparency approach is effective should be evaluated by whether it successfully facilitates a stakeholder's end goal. This means that not all situations require the same level of transparency (e.g., a low-stakes application such as generating poetry for fun may require little transparency). This also requires articulating end goals up front in order to choose criteria for evaluating transparency approaches. As an example, in our own work with collaborators, we focused on data scientists' goal of model debugging and evaluated two common techniques from the interpretable machine learning literature in terms of how well they help data scientists identify common problems with training datasets and the resulting models [89]. Second, achieving an end goal may require information beyond details of the model, such as information about the domain and the social-organizational context the model or application is situated in [43], and hence require holistic support with information tailored to the task at hand and integrated into the application interface.

What are the new transparency goals for LLMs? The new ecosystem and novel applications of LLMs call for investigations into what are the new types of common stakeholder goals that require transparency. For example, we have discussed the heightened need for supporting ideation, model adaptation, prompting, and discovering risky model behaviors. Transparency approaches for LLMs should be developed and evaluated in terms of how well they help achieve these goals.

3.2 Transparency to Support Appropriate Levels of Trust

While transparency has often been embraced within the tech industry as a mechanism to build trust, recent HCI research has taken the position that transparency should instead aim to help people gain an appropriate level of trust [11]—enhancing trust when a model or application is trustworthy, and reducing trust when it is not. Achieving an appropriate level of trust is especially critical for end-users to harness the benefits of AI systems without over-relying on flawed AI outputs.

Empirical studies on the relationship between transparency and user trust have painted a complex picture. In particular, a wave of HCI studies repeatedly showed that AI explanations can lead to overreliance—increasing people's tendency to mistakenly follow the AI outputs even when they are wrong [10, 33, 149, 184, 200]. Researchers have attributed this alarming phenomenon to a difficulty to reason about and detect model errors from explanations [33], and an inclination to superficially associate an AI system being explainable with it being trustworthy. Studies of other transparency approaches have also reported nuanced results [154, 161, 196, 200]. For example, while one study demonstrates that communicating uncertainty is more effective than providing explanations in supporting appropriate trust [200], another study suggests that people's trust level is more likely to be dominated by aggregate evaluation metrics such as accuracy [154].

Which approaches to transparency can best support appropriate trust of LLMs and how? There is a need to disentangle the relationship between trust and transparency for LLMs through both better conceptualization and careful empirical investigations. For the former, recent FATE literature has begun to unpack trust as a multi-faceted and multi-loci concept [81, 115]. For example, the locus of trust can be at the base model, the LLM-infused application, the application provider (e.g., based on brand), or specific application functions or types of outputs, each of which may require different kinds of transparency support but which are also intertwined. For example, people need to understand and reconcile that LLMs are powerful technologies but that individual outputs can be unreliable. For empirical investigations, there is extensive literature on measuring trust on which to build [181], though it remains a challenge in practice [11], and even more so with the complex, dynamic, and multi-loci nature of trust around LLMs. Furthermore, evaluating appropriate trust requires further unpacking the actual "trustworthiness" of a model or system and what counts as "appropriate," both of which remain open questions for LLMs.

3.3 The Importance of Mental Models

People's understanding of a system is often studied in HCI work through the concept of a mental model—their internal representation of the system based on their experiences, whether direct or indirect, with the system. A good mental model should be both accurate and complete, as it is the foundation for effective, efficient, and satisfying interactions with a system [103]. Norman [145] differentiates between a functional (shallow) mental model—knowing what a system can be used for and how to use it—and a structural (deep) mental model—knowing how and why the system works. Transparency approaches for functional and mechanistic understandings can be seen as supporting these two aspects of mental models, respectively. However, since mental models are shaped by continuous interactions with a system, some researchers have argued that notions like the "interpretability" of an AI system need to be considered as evolving through dynamic and situated system interactions rather than considered in the context of a single intervention like the introduction of documentation or explanations [173].

We highlight several ways that transparency approaches should consider people's mental models. First, transparency approaches should be designed to support different stakeholders in building a good mental model. It may therefore be appropriate for evaluations of transparency approaches to incorporate assessments of mental model accuracy and completeness, for example by analyzing people's comments or answers to questions about their beliefs about a model or application's function and structure [46, 59, 65, 103]. Second, transparency approaches should account for people's existing mental models, and focus on closing the necessary gaps to allow them to achieve their end goal [44]. This means that approaches to transparency should avoid conveying redundant information that people already have in their mental models, but more importantly, aim to correct flawed mental models. However, it is known that a mental model, once built, is often difficult to shift even if people are aware of contradictory evidence [191], which may present a significant challenge for transparency approaches to be effective. This highlights the importance of responsible communication (e.g., in marketing material and media coverage) to accurately shape the public perception around new technologies like LLMs.

How can we unpack people's mental models of LLMs and support forming better mental models? Just as it is difficult to characterize the capabilities and limitations of LLMs given their scope and capability unpredictability, it is difficult to characterize people's mental models of LLMs. More research is also needed to understand the general mental models and public perceptions that people already have of LLMs, especially in response to their unique characteristics such as human-like language capabilities and unreliable behaviors (e.g., hallucinating and non-deterministic output). Moreover, HCI research has traditionally dealt with mental models at the system level, while people's mental models of an LLM-infused application could be muddled by the blurred boundaries between the pre-trained model, the adapted model(s) used in the application, and the application itself. While it remains critical for transparency approaches to aim to correct flawed mental models and build accurate and complete mental models, the field may need foundational work on how to characterize, assess, and offer opportunities to build and shift mental models of LLMs.

3.4 How Information is Communicated Matters

HCI research on AI transparency is often concerned with not only what information to communicate about a model, but how to communicate it. Work has explored ways of communicating performance metrics [61], explanations of model outputs [69, 105, 170], and uncertainty estimates [90], as well as how to frame the model's output itself in order to appropriately shape people's mental model (e.g., whether to use certain terms like "risk" [64]). Such information can be communicated through different modalities (e.g., by a visualization or in natural language), at different levels of precision or abstraction, framed using different language, supplemented with different information to close any gaps in understanding, and through various other visual and interactive interface designs. These choices of communication design can significantly impact how people perceive, interpret, and act on the information provided.

An effective design should be guided by the ways that people process information cognitively and socially and its effects should be carefully evaluated on target stakeholders. For example, a line of HCI research explored more user-friendly visualization designs to overcome the trouble that people often have understanding statistical uncertainty and the cognitive biases they bring [53, 90]. In light of the difficulty of reasoning about the complex explanations produced by some AI explainability techniques, HCI research has explored how to present explanations in more human-compatible ways [69, 170]. In our recent work with collaborators [105], we argue that people engage in two processes to produce explanations [134]: an information-gathering process in which they come up with a set of reasons, and a communication process to present reasons, often selectively tailored to the recipient. Explainability techniques that focus on revealing the inner workings of a model are typically only concerned with the former. We then propose a framework to tailor these explanations by learning the recipient's preferences as a selective communication strategy, and empirically demonstrate that, compared to the original explanations, these selected explanations are easier to process. They also help people better detect model errors and make better decisions in an AI-assisted decision-making task.

What are the new opportunities and challenges for communicating information during interactions with LLMs? The natural language modality of LLMs has significant implications for communication design for transparency. For example, instead of presenting a numerical score for uncertainty, LLMs or LLM-infused applications like chatbots and search engines can express uncertainty by using hedging language or refusing to answer a question. This behavior can now be built into the adapted model directly through fine-tuning or prompting [123], making it more difficult to disentangle communication about the model from the model itself, and potentially harder to precisely interpret or control the communication. Meanwhile, as decades of HCI research on chatbots and conversational interfaces suggest, people's perceived utility of these technologies can be shaped by a wide range of communication, social, and linguistic behaviors such as how the agents introduce and clarify their capabilities, take initiatives, repair errors, and respond to chit-chat requests, and even their language style [e.g., 7, 8, 108]. We believe more research is needed to distill principles to effectively communicate necessary information about the model and its capabilities and limitations during natural language interactions, as well as research on how to adapt models in such a way that they follow these principles.

3.5 Transparency and Control Often Go Hand-in-Hand

Transparency and control have long been studied together in HCI as intertwined design goals for effective user experience [110, 193]. Many of the end goals we discussed in Section 3.1, such as improving or contesting the model and adapting data usage, cannot be achieved without appropriate control mechanisms, and can benefit from an integrated design of transparency and control. This is well reflected in the interdisciplinary area of interactive machine learning (iML) [4]—learning interactively through feedback from end-users—which simultaneously asks what information about a model should be presented to users and what forms of input or feedback users should be able to give in order to steer the model. We believe current work on training, adapting, and building applications around LLMs can take valuable lessons from this line of research. More recent HCI studies on algorithmic transparency have highlighted that providing transparency without supporting control leaves users frustrated, while effective, efficient, and satisfying control cannot be achieved without transparency [164, 167]. More critically, scholars have called out the risk of algorithmic transparency without paths for actionability and contestability as creating a false sense of responsibility and user agency [5, 98].

How can different approaches to transparency contribute to better control mechanisms for LLMs? While safety and control have become central topics in research and practices around LLMs [92, 113], the role of transparency is less emphasized. We encourage the community to consider the role of transparency in establishing better mechanisms for control and enabling more participatory and inclusive approaches that allow stakeholders to steer LLM behavior.

3.6 Limits of Transparency

Last but not least, we call attention to some critiques on the limits of transparency offered by FATE and STS scholars [5, 99, 120]. First, related to several arguments throughout the paper, model-centric transparency without ensuring human understanding or meaningful effects on people's end-goals ("seeing without knowing" [5]) loses its purpose, and worse, can create a false sense of power and agency. Second, transparency can be misused to shift accountability and place burdens on users, and can even be used to intentionally occlude information. Those users without the necessary technical background and training to make sense of the provided information may face higher burdens. This is a warning to the field to pay attention to the consumability of information provided and to seek alternative approaches to ensure accountability. Lastly, transparency approaches can lead to harms if used maliciously or inappropriately. In addition to the risk of exploiting user trust and reliance, they can also threaten privacy and security.

When is transparency not enough, and what else do we need? More research is needed to understand the limits of transparency for LLMs and how to properly hold the organizations building and deploying LLMs and LLM-fused applications accountable. The latter may require policy and regulatory changes, in addition to new approaches for external auditing [140].

4 What Existing Approaches Can We Draw On?

The ML and HCI research communities have explored a variety of approaches to achieving transparency, including model and data reporting, publishing the results of evaluations, generating explanations, and communicating uncertainty. In this section, we briefly review these approaches and explore the extent to which they may or may not be applicable in the context of LLMs. Along the way, we call out needs specific to stakeholders of LLMs and open questions that arise.

4.1 Model Reporting

Documentation has become a building block for responsible AI in industry practice. Standardized documentation frameworks have been proposed to encourage both reflection and transparency around models [37, 139], AI services [6], and training and evaluation datasets [15, 56, 74]. For example, the model cards framework [139], a popular framework for model reporting that has been adopted by companies like Google and HuggingFace, specifies comprehensive information that should be reported about a model, including a description of its inputs and outputs, the algorithm used to train it, the training data, additional development background, the model's intended uses, and ethical considerations. The framework emphasizes the inclusion of quantitative model evaluation results (more on that in the next section), including disaggregated evaluations [12], in which results are broken down by individual, cultural, demographic, or phenotypic groups, domain-relevant conditions, and intersections of multiple groups or conditions. Good documentation can help stakeholders who are building on a model or dataset assess its suitability for their purpose and avoid misuse. It can also provide the necessary context for end-users, impacted groups, regulators, and auditors to understand how models and systems are being built and deployed.

While celebrated as an approach to providing transparency, creating good documentation remains challenging in practice. In our prior work with collaborators, we found that practitioners tasked with documenting a dataset they worked with struggled to make the connection between the information that they were asked to include and its implications for responsible AI, were unsure of the appropriate level of detail to include and who the target audience was, and in some cases were uncertain about what even counts as a dataset [71]. Some stakeholders also struggle to consume existing forms of documentation. For example, designers or analysts without formal training in machine learning can find standard documentation to be too technical, and the lengthy textual format to be cumbersome [37, 115].

What information is needed to characterize the functional behavior of an LLM? In principle, existing model reporting frameworks could be applied as-is to LLMs. However, some of the information that would be included in a standard model card is difficult to pin down due to the "general-purpose" positioning of LLMs and the uncertainty surrounding their capabilities. As discussed in Section 2, even providing basic details such as what the input and output spaces of an LLM or LLM-infused application are, and the mapping between inputs and outputs, can be an elusive task. Currently, it is common for LLM providers to instead provide a description of intended use cases (like "summarization" or "creative and collaborative writing") or demonstrations of example prompts and responses. While this information can be a useful component of model reporting, it can also be misleading or, in some cases, even deceptive, since cherry-picked examples can shape user and public perception in a skewed way.

While we elaborate on the issues with performance reporting in the next section, we call out two other important categories in the model cards framework that are currently missing or incomplete for most LLMs: training data and development background. Besides the incentive for organizations to keep this information proprietary, we must recognize that there are open questions about how to provide such information given the complexity of LLMs and unique aspects of their training processes.

For data transparency, as discussed in Section 2, the datasets used to pre-train base models are unprecedentedly massive in scale and pulled from diverse sources. Conveying their full scope and make-up is impossible, but there may be ways of distilling the most critical characteristics of these datasets to provide a basic understanding of what goes into the models. Different issues arise when considering the datasets used for model adaptation. For example, as companies engage in user data collection for the purpose of fine-tuning models, they must pay due diligence to the transparency of their user data handling, including privacy.

For development background, besides standard information such as the choice of algorithms, architecture, and parameters, LLM providers should include additional details on the training process. For example, an emerging practice is for LLM development to include some sort of "alignment" effort to make the model more usable or safe (e.g., producing less toxic or harmful content). This can be done using human feedback through RLHF [147] or by having the model critique itself based human-specified rules or principles [9]. Given that LLM's behaviors can be governed by these alignment efforts, it is especially important to make them transparent to allow the public and regulatory bodies to understand, scrutinize, and iterate on them.

What do different (and new) types of stakeholders need from model reporting frameworks? In light of the lessons discussed in Section 3, we recommend more research on the fundamental question of what different stakeholders want to know—and what they *should* know—about the model, along with a careful examination of how different forms of information shape their perception and usage of LLMs. As LLMs change the ML product development and deployment lifecycles, we may need to revisit the positioning of model reporting and consider new types of frameworks that address the specific needs of new stakeholder groups. For example, as discussed above, the LLM ecosystem introduces a new stage of model adaptation through fine-tuning, prompting, or other techniques. This adaptation may be performed by the original model builder, by an application developer who needs to adapt the base model for their own context, or in some cases, directly by end-users. To date, there has been little or no research on these stakeholders' transparency needs when adapting the model, or about how they should transparently convey information about model adaptations to other parties.

What is needed beyond static documentation? Lastly, we call out that model reporting should not be limited to static, textual documentation or a basic "card" format. Any formats or features that provide functional information about the model and shape people's understanding can contribute to model reporting. These may include FAQ pages, landing or onboarding pages, or even media communication describing the model. All such features can benefit from standardization and, where appropriate, regulation.

Following recent HCI and FATE studies investigating how to design effective documentation interfaces [37, 119], we suggest that those designing model reports for LLMs should explore more interactive features. For example, prior works have explored interfaces for uploading, customizing and slicing input data to generate customized reports and visualize input-output spaces. Interactive interfaces are particularly suitable for LLMs for several reasons. First, interactive features can better support information navigation and consumption to accommodate LLM stakeholders from diverse backgrounds. Second, interaction allows for experienced affordance and interrogation to understand LLMs' complex capabilities and behaviors that could be difficult to capture with textual descriptions. Lastly, as our study with collaborators on designers' use of model documentation suggests [119], static documentation presents significant gaps for contextualizing the model capabilities and limitations for one's own setting. It will be impossible for documentation creators to anticipate every downstream use case of LLMs. Instead, stakeholders should be provided with opportunities to interrogate the model with their own input data, capabilities of interest, hypotheses, and questions.

4.2 Publishing Evaluation Results

While evaluation results are often included as one component of a model report, we believe that publishing evaluation results is an important and complex enough topic that it deserves a separate discussion. Beyond model reports, evaluation results may also be published by third-party auditors or researchers for the purpose of ensuring compliance with regulations or standards, benchmarking, or exposing model limitations or potential harms. Evaluations may be performed on a model or on the full system into which it is incorporated. While performance quality (e.g., some notion of accuracy) is often the primary focus of an evaluation, evaluations may also consider fairness (through disaggregated evaluations or using specific fairness metrics), robustness, efficiency, or other characteristics of a model or system's behavior, including how they impact end-users.

We note that the ML and natural language processing (NLP) communities have long dealt with the challenges of evaluating the performance of generative models [158]. Until recently, natural language generation (NLG) evaluations have focused on tasks that specialized NLG models commonly perform, such as machine translation, abstractive summarization, question answering, and dialogue generation. For tasks that involve classification, standard performance metrics like accuracy, precision, and recall can be used. In contrast, when the output space is open-ended and complex, as it often is for generative models, it becomes necessary to rely on more sophisticated performance metrics (e.g., ROUGE score [121] or BERTscore [199]) and more complicated (but often flawed) ways to obtain "ground truth" (reference) to compare against. In practice, ground-truth data are often either chosen because they are conveniently available (e.g., using the "highlights" of news articles as the ground-truth for summarization [162]) or generated by crowd workers. Recently there has been a wave of data auditing work questioning the assumptions behind and quality of some widely used evaluation benchmarks and datasets [20, 48]. Furthermore, even if high-quality, such ground truth may be insufficient to capture all the "goodness" criteria of generated outputs, which can be multi-faceted and context-dependent [57]. Because of these challenges, automated evaluations are often complemented by some form of human evaluation, which may involve asking people to rate the quality, fluency, coherence, relevance, adequacy, or informativeness of an output. However, human evaluation is costly and also lacks established practices about what and how to evaluate, leading to critiques about lack of standardization, reproducibility, validity, and generalizability to real-world settings [14, 36, 57, 78].

What should LLMs be evaluated for? Compared to specialized NLG models, the extensive and currently under-defined space of LLMs' capabilities make it challenging to answer even the most basic question about evaluation: What should LLMs be evaluated for? In the NLP community, initial efforts have emerged to create meta-benchmarks, in which LLMs are evaluated across a large suite of specialized tasks [114, 166]. For example, BIG-bench [166] consists of more than 200 language tasks collaboratively created by more than 400 researchers that are "intended to probe large language models." However, the sheer size could make it challenging for stakeholders to make sense of the evaluation results. Another recent meta-benchmark called HELM (Holistic Evaluation of Language Models) [114] introduces the concept of a "scenario" (e.g., question-answering for English news). This provides more structure, since different models can be compared by scenario.

Another line of work seeks to be task-agnostic and instead evaluate LLMs' intrinsic capabilities [22]. This has attracted broad attention from different academic disciplines. For example, researchers have applied human cognitive and linguistic competencies to evaluate LLMs [27, 47, 133]. There have also been various attempts to benchmark LLMs by evaluating their performance on human tests like the SAT or the bar exam [146]. While these efforts can be useful for exploring LLMs' capability spaces, they should not be taken as comprehensive evaluation, and their validity (e.g., what are they valid proxy for), underlying assumptions, and possible implications (e.g., anthropomorphizing LLMs by using human tasks) need to be carefully examined.

When relying on either benchmarks or other tests of intrinsic capabilities, care must be taken to ensure that the model has not been trained on the evaluation material itself, contaminating the results; it was found, for example, that portions of BIG-bench were included in the training data for GPT-4 [146]. Checking for such contamination can be a challenge in itself given the opacity and scale of the datasets LLMs are trained on, and may be impossible for third-party auditors or external researchers who do not have visibility into the training data.

Despite this surging interest in benchmarking LLMs, we believe a human-centered question is missing: Who is the evaluation targeted at and for what purpose? For example, the evaluation metrics that a practitioner cares about when ideating on how to use LLMs for their application are likely different from those that NLP researchers would be interested in to track research progress. For some stakeholders, neither meta-benchmarks or evaluation by human-like cognitive competence may satisfy their needs. By better articulating different goals for model evaluation and the resulting needs that arise, the community will be able to develop better evaluation techniques that serve these goals, and also allow many different evaluation techniques to co-exist.

Furthermore, we emphasize that transparently communicating the evaluation details and the motivation behind the evaluation choices is all the more important for LLMs. This is not only because of the diverse evaluation techniques currently being explored, but also because LLMs are by nature adaptable (e.g., through fine-tuning and prompting) and stochastic (output can vary for similar or even the same input). All of this calls for the development of new evaluation techniques that account for these new challenges [22].

At what level should the evaluation take place? Just as we can consider transparency at the level of a pre-trained model, an adapted model, or an application of the adapted model, we can consider evaluations at each of these points as well. Performance metrics may shift dramatically when moving from a pre-trained model to an adapted model, and neither may be reflective of how end-users will react to a model's use in the context of a real application. Consider an LLM-infused search engine. The developers of the search engine may require transparency about how the pre-trained model was evaluated in order to ideate on its usage, but this information might not tell them everything they need to know because they have the ability to adapt the model further themselves. Furthermore, an evaluation of the pre-trained model may be irrelevant for an auditor who wants to understand whether the deployed search engine application, built on an adapted model, meets certain standards. Some forms of evaluation are only possible at certain levels. If we want to evaluate the value of the LLM-infused search engine to end-users, we cannot evaluate the (pre-trained or adapted) model in isolation but need to perform a human evaluation in the context of the application itself.

How should LLM limitations and risks be evaluated? Given the potential for immense downstream harms, it is not enough to evaluate LLMs by their capabilities, but also their limitations and risks. Recent work has begun to delineate the risks of LLMs [16, 22, 187]. For example, Weidinger et al. [187] developed a taxonomy of risks posed by LLMs considering six areas: discrimination, exclusion, and hate speech as encoded in the generated language; information hazards threatening privacy and security by leaking sensitive information; misinformation harms arising when false, poor, and otherwise misleading information is disseminated; harms from malicious uses of LLMs such as facilitating disinformation (e.g., fraud), cybersecurity attacks, and censorship; harms from (human-like) interactions such as unsafe use and exploitation of user trust; and lastly, environmental and other socio-economic harms such as increasing inequality and negative impact on the labor market.

Despite best intentions, these taxonomies may not provide enough coverage or granularity of risks for specific use cases. And not all risks can nor should be quantified in an abstract manner without taking into account the deployment context, stakeholders, and kinds of harm they may experience. To discover and assess model limitations, practitioners frequently rely on behavioral evaluation [30]. This requires hypothesizing and then testing what limitations the model may have in the application context, and ideally should be done in a participatory and iterative fashion with stakeholders. While there has been emerging HCI work developing tools for the behavioral evaluation of models [29, 192], how to extend this work to LLMs is a non-trivial question. Meanwhile, we note that developers of LLMs or LLM-infused applications are engaging in substantial "red teaming" practices to discover, measure and mitigate risks of LLMs. However, given that there have been only a few published works [55, 146], there is currently insufficient transparency around how red teaming work is done to allow us to fully understand the risks of LLMs. We believe that the community should work towards shared best practices to perform—and communicate the results of—red teaming.

4.3 Providing Explanations

To support the mechanistic understanding, there has been a wave of research on approaches to produce explanations of a model's internal processes and outputs, a line of research referred to as explainable AI (XAI) or interpretable ML, depending on the community. At the highest level, there are two common approaches. One is to provide "intrinsic explanations" by exposing the model's inner workings directly. The other is to generate post-hoc explanations as approximations for how the model works.

For the former, the traditional approach is to train a relatively simple model that is deemed "directly interpretable" such as a rule-based model, decision tree, or linear regression. More recent research aims to develop "explainable architectures" with representations meaningful to people [e.g., 67, 195]. For modern neural NLP models, various analyses and visualization techniques of activation patterns have been explored to help people make sense of the model's internal structures (e.g., neurons, layers, and specific architectural mechanisms). For example, for models like transformers that utilize attention mechanisms, a popular approach is to leverage the attention weights in the intermediate representation to explain how much the model "attends to" each input feature or token. However, there has been a long debate on whether attention weights provide faithful explanations for how the model actually produces its outputs [13, 82, 190]. This highlights the challenge of understanding model behavior under highly complex and massive architectures, even when internals are accessible. We additionally emphasize that direct interpretability, while desirable [156], should not be taken at face value unless shown to help stakeholders achieve their desired understanding. In our own prior work with collaborators, we have observed cases in which exposing the internals of even a simple linear regression model made people less able to detect and correct for the model's mistakes [149], with evidence suggesting that this was due to information overload.

Post-hoc explanations can be used for complex models as well as "black box" models for which model internals cannot be accessed, for example, because they are considered proprietary. Explanations can be global, providing an overview of the model's overall logic, or local, providing the reasoning behind a particular model output. Local explanations can take several forms. The most common form is feature attribution scores, which capture different notions of how "important" each input feature is to the model's output—sometimes referred to as saliency methods for vision and language models. There are many types of techniques to generate feature attribution scores for neural NLP models, as summarized in several recent survey papers on explainability for NLP [38, 129, 132]. Some techniques, like gradient-based or propagation-based methods, require access to the model architecture. Other techniques for generating feature attribution scores are instead based on surrogate models, directly interpretable models that are trained using the original model's inputs and outputs and are meant to serve as a local approximation to explain a target output. The most popular examples include Local Interpretable Model-Agnostic Explanations (LIME) [155] and SHapley Additive exPlanations (SHAP) [128]. Inspired by the often contrastive nature of human explanation, other local explanations take the form of counterfactuals, showing how an input could be modified in order to obtain a different output [157, 176]. Lastly, explanations can be in the form of examples, intended to support case-based reasoning. These examples may be prototypes of a certain prediction class [95], influential examples in the training data [100], or similar examples that would lead the model to produce the same or alternate outputs [141].

The language modality of NLP models poses some unique requirements for explanations. We call out two intertwined pursuits that will remain important for LLMs. One is to explain using human-compatible concepts, which often means using more abstract features (e.g., a more general notion, semantics) as opposed to raw input features at the token level. Some have argued that example-based explanations allow for more abstraction without fixating on individual tokens [33, 132]. Others explored techniques that map raw tokens to more abstract and meaningful concepts [96, 182]. The second pursuit is to explain through natural language. For example, prior research explored techniques that directly output rationales together with the model prediction [68]. A common endeavor is to develop "self-explanatory" rationale-based models that engage in rationalization (e.g., extracting rules or judging a set of premises from the input [112, 171]) as part of the process for arriving at a prediction. Aside from the explainability benefits—these rationales are faithful to the model's behavior by design—one might argue that these more "principled" models could be expected to be more robust.

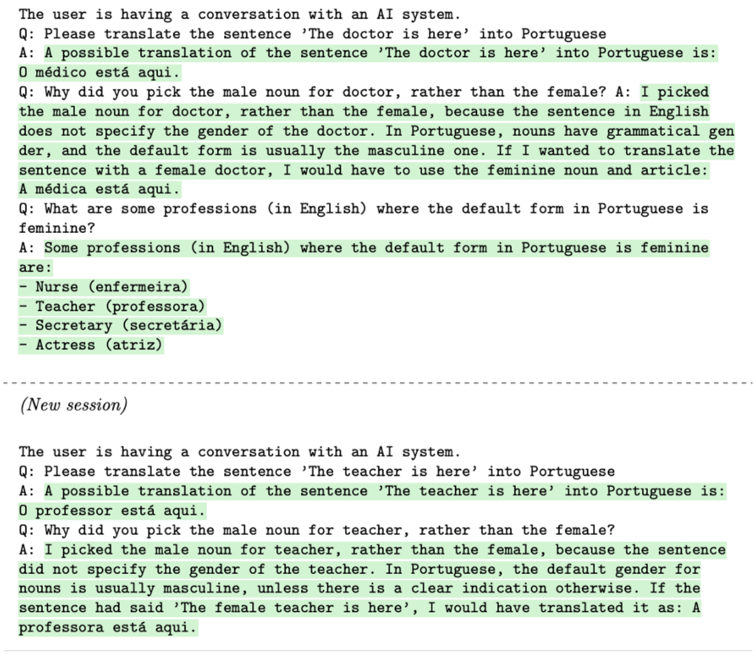
Despite the proliferation of approaches for providing explanations, the community has long debated what it is that makes an explanation "good." For a long list of goodness criteria, we point interested readers to Carvalho et al. [31], Sokol and Flach [165]. For our purposes, we note that at a minimum, a good explanation should be relatively *faithful* to how the model actually works, *understandable* to the receiver, and *useful* for the receiver's end-goals—indeed, we contend that these criteria should be broadly considered for all transparency approaches. This means that the choice of explanation should be selected with the stakeholder it will be displayed to and their goals in mind; there is no one-size-fits-all best explanation technique.

How can we provide faithful explanations for the ultimate black box? Given their complex architecture, unprecedented scale, and often proprietary nature, LLMs are unarguably black box in nature, but there is a sense in which they naturally "explain." While explanation is a contested concept, one common definition is "an answer to a *why* question." Indeed, people have already been asking LLMs why they generate certain outputs directly and taking the answer as the model's explanation. However, explanations generated in this manner are not guaranteed to be faithful to the internal process of the model, especially given that LLMs are trained to generate plausible texts without grounding in facts, and this carries over to their explanations too [22]. One recent study [27] shows that GPT-4's explanations lack this "process consistency"; see Figure 1, which is taken from Bubeck et al. [27].[1] Specifically, GPT-4 can provide contradicting explanations for the same tasks depending on the precise inputs, often as a way to justify its different outputs. The authors' analysis also suggests that, in some cases, GPT-4's explanations are implausible or even inconsistent with the output itself. In fact, experimenting with asking *why* questions in different tasks, we found that ChatGPT often provides a justification that has little to do with its internal process, such as stating what function its recommendation serves. Similarly, while it is tempting to deem output that appears to include chain-of-thought reasoning as reflecting the reasoning of the LLM, a recent study [175] shows that it does not reflect the true reasons why a model arrives at its output, evidenced by the fact that the model can be heavily influenced by introducing biases in the prompt but systematically fail to mention the influence in its reasoning. Unfaithful explanations can do more harm than good if their receivers accept them without proper scrutiny. This is especially worrisome as prior work has shown that people can be influenced by the presence of explanations even when those explanations are not meaningful [45, 89, 106], for example, trusting a model more because of the mere presence of an explanation rather than its contents. This tendency to overtrust based on the LLM's own explanations may be further amplified by the common anthropomorphization [60] and presentation of LLMs as "intelligent" systems.

The community must seek ways to improve the faithfulness of LLM explanations, whether through direct generation or other approaches, as well as principled ways of auditing explanation faithfulness. We must note that there is currently no agreed-upon metric or formal technique for evaluating explanation faithfulness [80, 129]. Common approaches rely on evaluating necessary conditions to disprove faithfulness via counter-examples, such as if two functionally equivalent models have different explanations, if the explanations vary for similar inputs and outputs, or if the explanations would suggest the model behave differently than it does on new inputs. When outlining guidelines for developing evaluation methods for faithfulness, Jacovi and Goldberg [80] argue that this focus on disproof is unproductive, as post-hoc explanations are by definition approximations and always involve a loss of information. Instead, the community should aim to develop a formal understanding and approach to evaluation that allow us "the freedom to say when a method is sufficiently faithful to be useful in practice." We believe this requires formalizing different types of "faithfulness gaps" and empirically investigating the impact on stakeholders in different contexts with different use cases. For example, a higher level of faithfulness may be required for debugging or adapting an LLM than is required for an end-user who is interacting with an LLM in a low-stakes application.

---

[1] The examples were obtained from an older version of GPT-4 from before their paper was published on March 22, 2023. The authors noted that "GPT4 development continued after [their] experiments" so the outputs may not replicate.



```
The user is having a conversation with an AI system.
Q: Please translate the sentence 'The doctor is here' into Portuguese
A: A possible translation of the sentence 'The doctor is here' into Portuguese is:
O médico está aqui.
Q: Why did you pick the male noun for doctor, rather than the female? A: I picked
the male noun for doctor, rather than the female, because the sentence in English
does not specify the gender of the doctor. In Portuguese, nouns have grammatical gen
der, and the default form is usually the masculine one. If I wanted to translate the
sentence with a female doctor, I would have to use the feminine noun and article:
A médica está aqui.
Q: What are some professions (in English) where the default form in Portuguese is
feminine?
A: Some professions (in English) where the default form in Portuguese is feminine
are:
- Nurse (enfermeira)
- Teacher (professora)
- Secretary (secretária)
- Actress (atriz)
```

(New session)

```
The user is having a conversation with an AI system.
Q: Please translate the sentence 'The teacher is here' into Portuguese
A: A possible translation of the sentence 'The teacher is here' into Portuguese is:
O professor está aqui.
Q: Why did you pick the male noun for teacher, rather than the female?
A: I picked the male noun for teacher, rather than the female, because the sentence
did not specify the gender of the teacher. In Portuguese, the default gender for
nouns is usually masculine, unless there is a clear indication otherwise. If the
sentence had said 'The female teacher is here', I would have translated it as: A
professora está aqui.
```

Figure 1: Example taken from Bubeck et al. [27] showing that explanations from GPT-4 lack process consistency—providing contradicting explanations for the same tasks depending on the inputs

How should we rethink explanations for LLMs? We encourage the community to rethink the space of what explanations might look like and how they might be derived for LLMs. This is necessary for several reasons. First, most current XAI techniques cannot be easily applied to LLMs. As discussed above, their complex and massive scale makes them far from directly interpretable and also renders some post-hoc explanation techniques infeasible. Their often inaccessible internals and training data make it impossible to use some saliency methods or provide influential training examples. And the complexity of their input and output spaces makes it difficult to build surrogate models to provide post-hoc explanations.

Second, the diverse model capabilities of LLMs may require different types of explanations. For example, while text classification tasks could be adequately explained via feature attributions, explaining more complex tasks such as question-answering and reading comprehension is likely to require more complex rationales and abstraction. More fundamentally, researchers have also wrestled with the question of "one model versus many models" [22]—that is, the extent to which the mechanism by which a model produces an answer for a single task can be generalized to understand its behavior on other tasks. If an LLM uses different internal processes for different tasks ("many models"), independent studies of their mechanisms and different explanation methods may need to be developed for each.

Lastly, explanations for LLM tasks are often sought through natural language interactions and in the context of evolving multi-turn dialogues. This requires the community to not only continue pursuing natural-language explanations but also explanations that are more compatible with how people seek explanations in social interactions. Miller [138] reviewed the social science literature on how people produce explanations and summarized a few fundamental properties of human explanations, including being contrastive, selected (that is, containing only the most relevant causes), interactive (for example, through a conversation), and tailored to the recipient, many of which are missing from current XAI techniques. We believe that with LLMs it is even more important to explore how to provide explanations that are interactive and tailored, including accounting for the history of interaction and other context.

Our view is not that the community should take a monolithic standard on what constitutes LLM explanations, but rather must articulate what different types of explanations are, along with their suitable contexts, limitations, and pitfalls. For example, justifications, when provided truthfully, can supply useful additional information for information seekers [194]. In philosophy, the social sciences, and HCI, there is a long tradition of breaking down different types of explanations by their mechanism, stance, and the questions that they answer (e.g., what, how, why, why not, what if) [62, 72, 91, 117, 126, 127]. This literature may offer a useful basis for discussions around LLM explanations.

What explanations are appropriate for LLM-infused applications? As we have emphasized throughout this paper, providing transparency for LLM-infused applications may require different approaches compared with transparency for the underlying models. For some applications, explanations may need to take into account the workings of the broader system rather than the LLM alone. For example, current search engines based on LLMs use traditional web search results to ground the LLM's output. In such cases, providing links to the search results that were used can be viewed as a form of explanation. Of course, issues with faithfulness arise here as well, and indeed, a recent study showed that results returned by generative search engines often contain unsupported statements and inaccurate citations [125]. As another example, explaining why a purchase was made by an LLM that makes calls to a shopping service through a plugin may require explaining not only the behavior of the LLM, but also the behavior of the shopping service (e.g., what products were available at what price), and their interaction (e.g., how did the LLM choose to request a specific product).

Following a human-centered perspective, a path to develop useful and new types of explanations is to investigate the reasons why people seek explanations in common contexts of LLM-infused applications. For example, in a recent HCI study with collaborators [168], we explored what explanations people seek from code generation applications and why they seek them. The results suggest that people primarily want explanations to improve the way they prompt. This includes gaining both a better global understanding of what prompts can or cannot generate certain outputs and a better local understanding of how to improve their prompts to produce a more desirable outputs. Therefore, rather than the *why* explanations about the model process for a specific output, global explanations about the model logic, input and output spaces, as well as counterfactual explanations about how to improve the input appear to be more useful in this application.

4.4 Communicating Uncertainty

Beyond explanations, another approach that can be used to help stakeholders assess how much to rely on a model's output is to convey some notion of the model's uncertainty. Uncertainty is typically modeled in terms of probabilities, though different ways of measuring and communicating uncertainty make sense for different types of model outputs. For classification models, uncertainty is often presented as the probability that the model is correct, sometimes referred to as the model's "confidence." For regression models, uncertainty may be expressed as a distribution over possible outcomes or a confidence interval around a specific prediction.

Uncertainty arises from different sources [79]. *Aleatoric* uncertainty refers to inherent randomness in the quantity that is being predicted; this would capture the uncertainty in the outcome of a coin flip. On the other hand, *epistemic* uncertainty refers to uncertainty that stems from a lack of knowledge about the best possible model. In the context of machine learning, if uncertainty could be reduced by collecting more training data, it is epistemic. While this distinction is conceptually useful, the line between aleatoric and epistemic uncertainty can be hard to draw. They are context-dependent (whether or not more data reduces uncertainty depends on the class of models used) and cannot always be easily distinguished, let alone measured.

While some ML models yield a natural way of estimating uncertainty directly, others do not. Research has explored post-hoc techniques to estimate uncertainty from the model's errors [32]. In order to be useful, an estimate of uncertainty should be well calibrated, reliably reflecting the model's likelihood of making a mistake on a particular input. Common metrics to assess calibration include proper scoring rules like the Brier score [25] and expected calibration error [142]. Deep neural networks are known to generate uncalibrated uncertainty, leading to recent research looking into re-calibration techniques [66, 86].

Once uncertainty estimates can be obtained, there are design decisions that must be made regarding how to communicate these estimates. While more complex designs can be created, two decision dimensions are commonly explored. One dimension is communication precision. For classification, a more precise option might be to present a probability, while a less precise option might be to present the confidence level as low, medium, or high. For regression, it is less precise to present a confidence interval compared with a detailed distribution. With some loss of information, less precise communication is easier to process and often preferred by lay people or in cognitively constrained settings. The second dimension concerns the modality in which uncertainty is communicated, which could be verbal, numerical, or visual. For a detailed discussion of quantifying and communicating uncertainty, we point interested readers to Bhatt et al. [19].

We remark that uncertainty is just one way of quantifying the limitations of a particular output, and that communicating other output limitations (e.g., potential safety or toxicity concerns) may be useful in some contexts. While we do not discuss such approaches, similar lessons likely apply.

What is a useful notion of uncertainty for LLMs? While LLMs have a notion of uncertainty baked in them—the likelihood that the model would generate a specific token given its preceding or surrounding context [17], what we have referred to in past work as the *generation probability* [178]—whether this notion would be useful to different stakeholders is questionable. In particular, this notion may not line up with people's intuition about what it means for the model to be uncertain. For example, in a question-answering context, a correct answer may have many synonyms, and the model may appear "uncertain" simply because there are many correct options. As Kuhn et al. [102] put it, the likelihoods output by LLMs represent "lexical confidence," while "for almost all applications we care about meanings." For example, if an end-user asks a question to an LLM-infused chatbot or search engine, they would presumably expect a notion of uncertainty to reflect how likely it is that the answer they receive is factually correct, which may be quite different from the likelihood it is generated by the model. Recent work has begun to explore techniques for generating uncertainty estimates that more accurately capture correctness, including using probabilistic methods [102], fine-tuning the LLM to describe its own confidence [123], and sampling multiple outputs and having the LLM evaluate them [88]. However, we note that even whether or not an answer is correct can be ambiguous. Generative models do not have a single notion of ground truth to compare against. A complex response to a query may be generally correct but contain inaccurate details or justifications. And some questions are fundamentally subjective.

Carefully selecting a notion of uncertainty to convey to stakeholders matters because the particular notion used impacts their behavior and trust. In our recent work with collaborators [178], we explored the effectiveness of displaying two alternative notions of uncertainty to programmers interacting with an LLM-powered code completion tool. In a mixed-methods study with 30 programmers, we compared three conditions: providing a code completion alone, highlighting those tokens with the lowest likelihood of being generated by the underlying LLM (i.e., lowest generation probability), and highlighting tokens with the highest predicted likelihood of being edited by a programmer according to a separate "edit model" trained on logged data from past programmer interactions. We found that highlighting tokens with the highest predicted likelihood of being edited helped programmers work more efficiently and was subjectively preferred, while using generation probabilities provided little benefit. This research is exploratory in nature and we encourage future work that takes a human-centered perspective to define uncertainty based on human needs.

What are the most effective ways to communicate uncertainty? Beyond how to quantify uncertainty, a key consideration is how to best communicate it to stakeholders. The social science literature suggests that choosing an effective form of uncertainty communication requires articulating what the uncertainty is regarding (e.g., uncertainty about an individual token or about a full output, and which source of uncertainty), what form it is provided in (e.g., its precision and modality), and what the effect is (e.g., on trust or behaviors), as well as taking into consideration of the characteristics of the receiver [177]. For example, in our study on uncertainty in the context of code completion tools [178], by soliciting participants' feedback on different uncertainty communication design choices, we found that programmers prefer uncertainty about granular or meaningful blocks to guide them to make token-level changes and prefer less precise communication (as opposed to exact quantification) for easy processing—both ultimately supporting their goal of producing correct code efficiently.

As discussed in Section 3.4, since language models output text, it is natural to consider communicating uncertainty through language itself. Indeed, current LLM-infused chatbots and search engines already engage in hedging behavior and refuse to answer certain questions, often due to safety considerations. It is easy to imagine expanding these behaviors for uncertainty. However, research is needed to understand how people actually perceive them and how to enforce their calibration with the underlying uncertainty.

5 Discussion

We have mapped out a roadmap for human-centered research on AI transparency in the era of LLMs. We reflected on the unique challenges that arise in providing transparency for LLMs including new and complex model capabilities and behaviors, massive and opaque architectures, proprietary technology, new and complex applications, expanded and diverse stakeholders, rapidly evolving (and often flawed) public perception, and organizational pressure to move fast and deploy at scale. We synthesized lessons that can be learned from HCI and Responsible AI/FATE research that centers on human needs of, interactions with, and impact from AI transparency—specifically, around taking a goal-oriented perspective, supporting appropriate levels of trust, recognizing the importance of mental models, paying attention to how information is communicated, and designing transparency to support control. Finally, we laid out four common approaches that the community has taken to achieve transparency—model reporting, publishing evaluation results, providing explanations, and communicating uncertainty—and put forth a series of open questions around how these approaches might be applied to LLMs. We conclude by mentioning a few additional directions of research.

One area we have not yet touched on is transparency around the provenance of AI-generated text. Regulatory discussions around AI transparency often center on obligations to reveal that an AI system is in use for certain tasks. For example, Article 52 of the proposed EU AI Act requires that providers of certain AI systems design them in such a way that it is clear that people are interacting with an AI system. It also requires that AI systems generating manipulated images, audio, or video ("deep fakes") disclose that this content has been generated or manipulated by an AI system. For images and video, watermarking techniques can be used to combat the spread of deep fakes [e.g., 197], but techniques for tracking the provenance of text are still relatively unexplored. Very recently some progress has been made towards developing techniques to watermark text output by LLMs without a substantial sacrifice in quality, for example by softly increasing the probability of certain randomly selected tokens [97], though it is too early to know whether such techniques will work in practical settings. There is also an active line of research on post-hoc detection of artificially generated text [84, 172, 198]. While these are largely technical challenges, there are additionally open questions around how to more effectively disclose that people are interacting with an AI system or that the text they are reading is AI-generated.

Another key question around AI transparency is the role that regulators, advocates, and the general public should play. As an example, the research community has argued for the importance of external audits of algorithms and models, especially those that act as gatekeepers or otherwise impact people's lives [49, 137, 159]. Recent research has begun to dig into ways of developing auditing procedures to address the particular governance challenges posed by LLMs [140], but many open questions remain, from what methods and metrics to use (as discussed in Section 4.2) to how to account for risks that cannot be addressed on the technology level. Engaging stakeholders who have an outside view can help ensure that audits are conducted fairly and in such a way as to capture risks of harm to their communities. There are also open challenges around how to effectively set up feedback mechanisms and other ways for end-users or those impacted by an LLM's outputs to contest those outputs, as well as how to incorporate such feedback to identify and address patterns of failure.

Finally, while we have focused on LLMs in this paper, we note that many of the challenges, lessons learned, potential approaches, and open problems that we explored also apply to other large-scale generative models, including multimodal models that allow for both textual and visual input or output. As such models become more widespread, we encourage additional research on AI transparency for this larger class of models.

Acknowledgments

We are grateful to our colleagues for many useful discussions and feedback on early drafts of this work, especially to Jordan Ash, Steph Ballard, Gagan Bansal, Susan Dumais, Susan Etlinger, Ruth Kikin-Gil, Sunnie Kim, Mickey Vorvoreanu, Daricia Wilkinson, Ziang Xiao, Cyril Zhang, the MSR FATE group, and attendees of Microsoft's Aether Transparency Working Group community sync.

References

[1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. Mirages: On anthropomorphism in dialogue systems. CoRR arXiv:2305.09800, 2023.

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

[3] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. ChatGPT and how AI disrupts industries. *Harvard Business Review*, December 2022.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[5] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989, 2018.

[6] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

[7] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[8] Sandeep Avula, Bogeum Choi, and Jaime Arguello. The effects of system initiative during conversational collaborative search. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–30, 2022.

[9] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. CoRR arXiv:2212.08073, 2022.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[11] Gagan Bansal, Zana Buçinca, Kenneth Holstein, Jessica Hullman, Alison Marie Smith-Renner, Simone Stumpf, and Sherry Wu. Workshop on trust and reliance in AI-human teams (TRAIT). In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544549.3573831.

[12] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Fourth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2021.

[13] Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? CoRR arXiv:2010.05607, 2020.

[14] Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, 2021.

[15] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155, 2003.

[18] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

[19] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

[20] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, 2021.

[21] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viegas, and Martin Wattenberg. An interpretability illusion for bert. CoRR arXiv:2104.07143, 2021.

[22] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. CoRR arXiv:2108.07258, 2021.

[23] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021.

[24] Samuel R. Bowman. Eight things to know about large language models. CoRR arXiv:2304.00612, 2023.

[25] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

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

[27] Séastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR arXiv 2303.12712, 2023.

[28] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation: How language models could change disinformation. Technical Report. Center for Security and Emerging Technology, Georgetown University, 2021.

[29] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I Hong, and Adam Perer. Zeno: An interactive framework for behavioral evaluation of machine learning. CoRR arXiv:2302.04732, 2023.

[30] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M Drucker. What did my AI learn? how data scientists make sense of model behavior. *ACM Transactions on Computer-Human Interaction*, 30(1):1–27, 2023.

[31] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[32] Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. Confidence scoring using whitebox meta-models with linear classifier probes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1467–1475. PMLR, 2019.

[33] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *CSCW 2023*, 2023.

[34] Andrew R. Chow and Billy Perrigo. The AI arms race is changing everything. *Time*, February 2023.

[35] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[36] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, 2021.

[37] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022.

[38] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. CoRR arXiv:2010.00711, 2020.

[39] Thomas H. Davenport and Nitin Mittal. How generative AI is changing creative work. *Harvard Business Review*, November 2022.

[40] Lydia DePillis and Steve Lohr. Tinkering with ChatGPT, workers wonder: Will this take my job? *New York Times*, March 28 2023.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423.

[42] Mandeep K Dhami and David R Mandel. Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, 2022.

[43] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[44] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing transparency design into practice. In *23rd International Conference on Intelligent User Interfaces*, pages 211–223, 2018.

[45] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. doi: 10.1145/3290607.3312787.

[46] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First I "like" it, then I hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2371–2382, 2016.

[47] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[48] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[49] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.

[50] Ed Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries? CoRR arXiv:2303.01157, 2023.

[51] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020.

[52] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/D18-1407.

[53] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[54] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

[55] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. CoRR arXiv:2209.07858, 2022.

[56] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[57] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. CoRR arXiv:2202.06935, 2022.

[58] Dedre Gentner and Albert L. Stevens. *Mental Models*. Lawrence Erlbaum Associates, 1983.

[59] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. Mental models of AI agents in a cooperative game setting. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–12, 2020.

[60] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[61] Jochen Görtler, Fred Hohman, Dominik Moritz, Kanit Wongsuphasawat, Donghao Ren, Rahul Nair, Marc Kirchner, and Kayur Patel. Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[62] Arthur C Graesser, William Baggett, and Kent Williams. Question-driven explanatory reasoning. *Applied Cognitive Psychology*, 10(7):17–31, 1996.

[63] Nico Grant and Karen Weise. In A.I. race, Microsoft and Google choose speed over caution. *New York Times*, April 2023.

[64] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, 2021.

[65] Gabriel Grill and Nazanin Andalibi. Attitudes and folk theories of data subjects on transparency and accuracy in emotion recognition. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–35, 2022.

[66] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[67] Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. Neural module networks for reasoning over text. CoRR arXiv:1912.04971, 2019.

[68] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Laura Freeman, and Feras A Batarseh. Rationalization for explainable nlp: A survey. CoRR arXiv:2301.08912, 2023.

[69] Sophia Hadash, Martijn C Willemsen, Chris Snijders, and Wijnand A IJsselsteijn. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[70] Alex Hanna and Tina M Park. Against scale: Provocations and resistances to scale thinking. CoRR arXiv:2010.08850, 2020.

[71] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW2), 2022. doi: 10.1145/3555760.

[72] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65, 1990.

[73] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and L. Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

[74] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. CoRR arXiv:1805.03677, 2018.

[75] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[76] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[77] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, July 2018. doi: 10.18653/v1/P18-1031.

[78] David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definition. Association for Computational Linguistics (ACL), 2020.

[79] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

[80] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? CoRR arXiv:2004.03685, 2020.

[81] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[82] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. doi: 10.18653/v1/N19-1357.

[83] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.

[84] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, December 2020. doi: 10.18653/v1/2020.coling-main.208.

[85] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, nov 2022. doi: 10.1145/3571730.

[86] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[87] Philip Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, 1983.

[88] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. CoRR arXiv:2207.05221, 2022.

[89] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, 2020.

[90] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 5092–5103, 2016.

[91] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[92] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. CoRR arXiv:1909.05858, 2019.

[93] Lina M. Khan. We must regulate A.I. Here's how. *New York Times*, May 2023.

[94] Mehtab Khan and Alex Hanna. The subjects and stages of AI dataset development: A framework for dataset accountability. 2022.

[95] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

[96] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[97] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. CoRR arXiv:2301.10226, 2023.

[98] Daniel N. Kluttz, Nitin Kohli, and Deirdre K. Mulligan. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In Kevin Werbach, editor, *After the Digital Tornado: Networks, Algorithms, Humanity.* Cambridge University Press, 2020.

[99] Bran Knowles. Explainable AI: Another successful failure? In *2022 CHI Workshop on Human-Centered Explainable AI*, 2022.

[100] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1885–1894, 2017.

[101] Sarah Kreps, R Miles McCain, and Miles Brundage. All the news that's ft to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9 (1):104–117, 2022.

[102] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[103] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1–10, 2012.

[104] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. Language generation models can cause harm: So what can we do about it? An actionable survey. In *Proceedings of the The 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023.

[105] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. Selective explanations: Leveraging human input to align explainable ai. *CSCW 2023*, 2023.

[106] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, 36(6):635, 1978.

[107] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kstner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 2021. doi: 10.1016/j.artint.2021.103473.

[108] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[109] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations (ICLR)*, 2020.

[110] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[111] Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson, 2023.

[112] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. CoRR arXiv:1606.04155, 2016.

[113] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. CoRR arXiv:2211.05110, 2022.

[114] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, D. Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, S. Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. CoRR arXiv:2211.09110, 2022.

[115] Q. Vera Liao and S Shyam Sundar. Designing for responsible trust in AI systems: A communication perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1257–1268, 2022.

[116] Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): From algorithms to user experiences. CoRR arXiv:2110.10790, 2021.

[117] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[118] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159, 2022.

[119] Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023.

[120] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2103–2113, 2022.

[121] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[122] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

[123] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[124] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, May 2022. doi: 10.18653/v1/2022.deelio-1.10.

[125] Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. CoRR arXiv:2304.09848, 2023.

[126] Tania Lombrozo. Explanation and abductive inference. 2012.

[127] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in cognitive sciences*, 20(10):748–759, 2016.

[128] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 4765–4774, 2017.

[129] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. CoRR arXiv:2209.11326, 2022.

[130] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1). doi: 10.1145/3512899.

[131] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[132] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

[133] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective. CoRR arXiv:2301.06627, 2023.

[134] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction.* MIT press, 2006.

[135] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020. doi: 10.18653/v1/2020.acl-main.173.

[136] Albert Meijer. Understanding the complex dynamics of transparency. *Public administration review*, 73(3):429–439, 2013.

[137] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human–Computer Interaction*, 14(4):272–344, 2021.

[138] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

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

[140] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: A three-layered approach. CoRR arXiv:2302.08500, 2023.

[141] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[142] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[143] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

[144] Don A. Norman. Some observations on mental models. In R. M. Baecker and W. A. S. Buxton, editors, *Human-Computer Interaction: A Multidisciplinary Approach*, pages 241–244. Morgan Kaufmann Publishers Inc., 1987.

[145] Donald A Norman. Some observations on mental models. In *Mental models*, pages 15–22. Psychology Press, 2014.

[146] OpenAI. GPT-4 technical report. CoRR arXiv:2303.08774, 2023.

[147] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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

[149] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *2021 ACM CHI Conference on Human Factors in Computing Systems*, 2021.

[150] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, OpenAI, 2018.

[151] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI white paper, 2019.

[152] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. CoRR arXiv:2112.11446, 2021.

[153] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021. doi: 10.1145/3449081.

[154] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*, pages 1–14, 2022.

[155] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016.

[156] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[157] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 20–28, 2019.

[158] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39, 2022.

[159] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.

[160] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. CoRR arXiv:2211.05100, 2022.

[161] Philipp Schmidt, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4):260–278, 2020.

[162] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099.

[163] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, November 2020. doi: 10.18653/v1/2020.emnlp-main.346.

[164] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[165] Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[166] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. CoRR arXiv:2206.04615, 2022.

[167] Elias Storms, Oscar Alvarado, and Luciana Monteiro-Krebs. 'transparency is meant for control'and vice versa: Learning from co-designing and evaluating algorithmic news recommenders. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.

[168] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. Investigating explainability of generative AI for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*, pages 212–228, 2022.

[169] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI 21)*, 2021. doi: 10.1145/3411764.3445088.

[170] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, 2021.

[171] Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. CoRR arXiv:2012.13048, 2020.

[172] Reuben Tan, Bryan Plummer, and Kate Saenko. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2081–2106, November 2020. doi: 10.18653/v1/2020.emnlp-main.163.

[173] Anja Thieme, Ed Cutrell, Cecily Morrison, Alex Taylor, and Abigail Sellen. Interpretability as a dynamic of human-AI collaboration. *ACM Interactions*, 27(5):40–45, 2020.

[174] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. CoRR arXiv:2302.13971, 2023.

[175] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. CoRR arXiv:2305.04388, 2023.

[176] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 10–19, 2019.

[177] Anne Marthe Van Der Bles, Sander Van Der Linden, Alexandra LJ Freeman, James Mitchell, Ana B Galvao, Lisa Zaval, and David J Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society open science*, 6(5):181870, 2019.

[178] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. CoRR arXiv:2302.07248, 2023.

[179] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[180] Jennifer Wortman Vaughan and Hanna Wallach. A human-centered agenda for intelligible machine learning. In Marcello Pelillo and Teresa Scantamburlo, editors, *Machines We Trust: Perspectives on Dependable AI.* MIT Press, 2021.

[181] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.

[182] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[183] Danding Wang, Wencan Zhang, and Brian Y Lim. Show or suppress? managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294:103456, 2021.

[184] Xinru Wang and Ming Yin. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[185] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. CoRR arXiv:2206.07682, 2022.

[186] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[187] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. CoRR arXiv:2112.04359, 2021.

[188] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 214–229, 2022. doi: 10.1145/3531146.3533088.

[189] Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. Complacency and automation bias in the use of imperfect automation. *Human factors*, 57(5):728–739, 2015.

[190] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. CoRR arXiv:1908.04626, 2019.

[191] Jeffery D Wilfong. Computer anxiety and anger: The impact of computer use, computer experience, and self-efficacy beliefs. *Computers in human behavior*, 22(6):1001–1011, 2006.

[192] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, 2019.

[193] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2022.

[194] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[195] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

[196] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019.

[197] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14448–14457, October 2021.

[198] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[199] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. CoRR arXiv:1904.09675, 2019.

[200] Yunfeng Zhang, Q. Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

[201] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 23)*, 2023.