Human-Centered LLM

44

Giulia Di Fede, Lina Alrabie, and Salvatore Andolina

Contents

G. Di Fede · L. Alrabie · S. Andolina (✉)
DEIB, Politecnico di Milano, Milan, Italy
e-mail: giulia.difede@polimi.it; lina.alrabie@polimi.it; salvatore.andolina@polimi.it

W. Xu (ed.), *Handbook of Human-Centered Artificial Intelligence*,
https://doi.org/10.1007/978-981-95-3658-0_49

Abstract

This chapter discusses the concept of human-centered large language models (LLMs). It explores both the opportunities and challenges involved in designing LLMs that are aligned with human values, highlighting the significance of human-centricity and reflecting on key considerations such as inclusivity, transparency, controllability, and sustainability. The chapter examines key challenges and how they arise across the LLM design lifecycle, advocating for an integrated approach that considers not only the development of the underlying models but also the design of the interfaces and interaction methods through which people engage with them. It provides an overview of current developments in the field and discusses key design principles and case studies of effective human-AI interaction. The chapter concludes by highlighting important trends and prospects in various sectors.

Keywords

Human-centered AI · Large language models · HCI · HCAI

1 Introduction

1.1 Overview of LLMs

Large language models (LLMs) represent a revolutionary advancement in the field of artificial intelligence. They are trained on vast datasets and designed to understand and generate human language with remarkable fluency. The release of ChatGPT on November 30, 2022, marked a pivotal moment in AI history, making large language models widely accessible through an intuitive, chat-based interface. This breakthrough significantly raised public awareness of LLMs' potential, as these models demonstrated impressive emerging capabilities including answering questions, translating languages, summarizing information, and generating original content. At the core of these interactions is the concept of a *prompt*—a user-provided input that guides the AI's response. For example, a simple request like "Explain quantum mechanics in laypeople's terms" allows users to instantly access complex knowledge in an easy-to-understand format, a task that previously required extensive research or expert consultation. Leading examples of these AI models include proprietary systems like OpenAI's GPT series, Anthropic's Claude, and Google's Gemini, as well as open-source models such as Meta's LLaMA, Mistral, and Deepseek, which power applications such as chatbots, virtual assistants, and tools for education (Kasneci et al., 2023), creativity (Di Fede et al., 2022), and business automation

(Kumar et al., 2024b; Dalsaniya & Patel, 2022). However, it soon became evident that, despite their impressive abilities, LLMs also faced notable issues such as generating inaccurate information, reflecting biases present in their training data, and encountering limitations in logical reasoning. As these systems become increasingly integrated into society, questions about their deployment, oversight, and social impact have taken center stage in discussions on AI ethics (Bellogín et al., 2024).

Recently, research communities have increasingly turned their attention to the effects of AI on human decision-making and its broader impact on humans (Andolina & Konstan, 2023). This shift has given rise to new movements and research paradigms, most notably human-centered AI (HCAI) (Ozmen Garibay et al., 2023). By prioritizing the well-being of humans and the environment they live in, the HCAI approach challenges us to redefine success in AI, not solely by technical performance but by its ability to support ethical and meaningful interactions. Given the recent explosion of LLMs and their increasing impact on society, it is now crucial to adapt these principles specifically to LLMs, redefining what it means to design and develop truly human-centered language models.

1.2 Importance of Human-Centricity in LLMs

As LLMs become increasingly integrated into everyday life—from education and healthcare to legal advice and social platforms—it is imperative to ensure the effectiveness and safety of these systems. Human-centricity in LLMs involves grounding the design and deployment of these systems in human needs, values, and expectations. This means considering not only what these systems should be capable of, but also how they would influence those who interact with them. A growing body of research in human-computer interaction (HCI) warns that designing LLMs without human-centricity as the main focus can lead to harmful consequences, including loss of trust and the reinforcement of social inequities.

Human-centered LLMs aim to integrate human-centric principles throughout their lifecycle. In line with HCAI principles, design should prioritize not just performance, but algorithmic transparency, fairness, and appropriate trust (Amershi et al., 2019; Ozmen Garibay et al., 2023). However, in practice, many LLM deployments have failed to meet these criteria, often with serious consequences.

The AI Incident Database (McGregor, 2021) documents over a hundred failures, illustrating how LLMs can perpetuate harm and misinformation when deployed without a human-centered focus. For instance, biased outputs from GPT-3—such as associating certain demographic descriptors with violent content—demonstrate how training data can reflect and amplify harmful stereotypes. In a particularly tragic case, an LLM-powered chatbot was reported to have encouraged a vulnerable user toward self-harm during emotionally charged conversations. Misinformation also presents a serious concern; Meta's Galactica, designed to assist with scientific writing, was quickly withdrawn after it began producing authoritative-sounding but factually incorrect and sometimes offensive outputs. Meanwhile, breaches of user trust have occurred in more subtle forms: a mental health platform used GPT-3 to co-author responses to individuals in crisis without informing them, raising ethical concerns around transparency, consent, and responsible deployment. Another concern regards *sycophancy* in language models, referring to their tendency to respond with excessive agreement, praise, or affirmation. A notable example where this tendency was particularly evident to the public is the GPT-4o update in late April 2025, after which users observed that ChatGPT had begun responding in an overly validating and agreeable manner, even to questionable or dangerous inputs. OpenAI subsequently rolled back the update, citing the need for more balanced behavior. These incidents underscore the risks of LLMs when deployed without a strong human-centric foundation. The HCI community has long advocated for principles that promote human-centricity and demonstrated how these could mitigate some of the current challenges faced by LLMs as well as expose additional hidden challenges emerging when humans interact with these systems (Lee et al., 2025).

We argue that human-centricity is not an optional feature in LLM development, but a societal necessity that is crucial to building systems that not only demonstrate technical excellence but also respect and promote human values. By embedding human-centric principles at every stage, traditional LLMs can evolve into tools that truly serve and empower people; that is what we call *human-centered LLMs*.

1.3 Scope and Objectives of This Chapter

This chapter seeks to provide an overview of the challenges, opportunities, and future directions involved in the design and implementation of human-centered LLMs. It begins by highlighting the significance of human-centricity in LLMs and the key considerations that influence their development. The second section offers a background on data sources, training processes, and the role of LLMs in modern AI systems. It also discusses the fundamental principles of human-centered design (HCD) for LLMs. The following section delves into the main challenges associated with human-centered LLM, such as human value alignment, bias and fairness, transparency, privacy, and accessibility. This section also provides insights regarding addressing and mitigating such challenges. In Sect. 4, a structured design framework will be presented to illustrate the phases of the human-centered LLM design process. Additionally, this section covers related concepts, including data collection and curation, AI explainability, user interactions and controllability, accessible and inclusive LLMs, and evaluation with human subjects. To better understand real-world applications of human-centric principles in LLM development, Sect. 5 presents two case studies that showcase effective user interactions with LLMs. Section 6 highlights emerging trends, emphasizing the scalability, human-AI alignment, and sector-specific prospects of human-centered LLMs. Finally, the last section concludes with a summary of the chapter's highlights and key insights, reinforcing the importance of ethical, transparent, and sustainable LLM development.

2 Background

2.1 Data Sources and Training

In the twenty-first century, data has emerged as the most valuable resource, often referred to as the gold of our time. LLMs derive their versatility from learning patterns, structures, and knowledge from massive and diverse datasets. Central to this learning process are the concepts of lexicon and corpus, which respectively provide the semantic foundations and real-world context for language understanding and generation. This enables them not only to generate fluent text but also to exhibit complex capabilities such as reasoning, decision-making, and multistep task execution. However, these strengths are tightly linked to the quality, diversity, and representativeness of the data they consume. For this reason, developing human-centered LLMs requires a thoughtful approach to data selection and training methodologies to ensure their behavior aligns with human values.

The training of LLMs typically involves two distinct stages, each relying on different kinds of data and objectives: *pre-training* and *post-training*.

Pre-training involves exposing the model to large, diverse datasets drawn from sources such as Common Crawl (web text), Wikipedia (encyclopedic content), Web-Text (user-curated data), Project Gutenberg (books), and arXiv (academic literature). These datasets enable the model to learn linguistic structures and general knowledge across a wide range of domains. Two key processes guide pre-training (Zhao et al., 2023; Zha et al., 2025):

- *Data Preprocessing* cleans the raw data by removing noise, low-quality content, and sensitive information (e.g., personally identifiable information) and prepares the text for tokenization.
- *Data Scheduling* organizes how different datasets are mixed and presented during training (e.g., their order and proportions), ensuring balanced and systematic exposure (Xie et al., 2023).

Post-training tailors the model for real-world use by improving its ability to follow instructions and align with human values. This phase generally includes two main stages: instruction tuning and alignment tuning (Zhao et al., 2023).

- *Instruction Tuning* (also known as supervised fine-tuning) enhances the model's responsiveness to human instructions by training it on input-output pairs. Human-curated datasets like Natural Instructions (Mishra et al., 2022) offer high-quality examples but are limited in scale. In contrast, synthetic datasets such as Self-Instruct (Wang et al., 2023a) and Baize (Xu et al., 2023) provide broader coverage but risk perpetuating errors or biases from the generating models.
- *Alignment Tuning* is meant to improve the model's capability of generating outputs that reflect human values and avoid unintended harms, such as biased language, misinformation, or manipulative behavior (Wang et al., 2023b; Kenton et al., 2021). Misaligned models pose ethical risks that go beyond task performance, potentially undermining trust and safety in AI systems.

A widely used method for alignment tuning is reinforcement learning from human feedback (RLHF) (Ziegler et al., 2020). This approach incorporates human judgments into the training process to guide the model toward more preferred behaviors. It typically begins with supervised fine-tuning, where the model is trained on high-quality examples to establish a foundation of desirable responses. Next, human evaluators compare and rank different model outputs. These rankings are used to train a reward model that predicts which responses are more aligned with human preferences. Finally, the model is further optimized using reinforcement learning—commonly with methods like proximal policy optimization (Schulman et al., 2017)—so it learns to produce responses that maximize the predicted reward (Ouyang et al., 2022). While RLHF has improved the alignment of popular models, it faces scalability challenges, as high-quality human feedback is both costly and subjective. These limitations highlight the need for continued research into more scalable, robust, and principled approaches to alignment. To address these limitations, newer techniques are emerging (Whitmore et al., 2024; Lee et al., 2024), but more research is needed to understand how human-centric principles could support this important phase.

2.2 Role of LLMs in Modern AI Systems

LLMs are revolutionizing the way work is accomplished in various sectors (Butler et al., 2024). In domains such as design, research, education, and innovation, LLMs have become powerful tools able to assist with ideation, accelerate iteration cycles, and support decision-making processes. For example, LLMs have been integrated into creative workflows to enhance brainstorming, refine prototypes, and streamline product design (Bilgram & Laarmann, 2023; Di Fede et al., 2022).

Similarly, in knowledge-intensive environments—such as scientific research, journalism, legal analysis, and healthcare—LLMs serve as accelerators for information synthesis and reasoning. By parsing vast corpora of unstructured data, LLMs assist in generating concise summaries, comparing sources, or even hypothesizing new relationships. In clinical and biomedical contexts, they are widely used to streamline documentation, reduce the burden of post-visit note-taking, and help mitigate clinician burnout. One of the most impactful applications in this sense is AI-powered ambient dictation systems, or AI scribes, which transcribe and summarize clinician–patient interactions in real time (Tierney et al., 2025). Beyond documentation, LLMs are enhancing clinical decision support, diagnostics, and research. They also play an emerging role in drug discovery and personalized medicine by helping identify therapeutic candidates and predict treatment outcomes (Clusmann et al., 2023; Zheng et al., 2024).

In all these fields and more, LLMs have been integrated in diverse and innovative ways, enabling new possibilities for streamlining progress. The next step in the evolution of AI-driven solutions is the emergence of AI agents, which leverage LLMs as their core engine to autonomously interact, make decisions, and drive actions across a wide range of applications. In particular, agentic approaches put an LLM in charge of directing the flow and logic of tasks by dynamically adapting to new challenges through reasoning while coordinating actions and leveraging available tools to achieve specific goals. AI agents powered by LLMs have already been applied to a variety of domains, including social sciences, natural sciences, and engineering, demonstrating their versatility and potential across diverse fields (Wang et al., 2024).

While LLM-based agents already represent a significant advancement in autonomous AI systems, their capabilities are being further advanced with the introduction of LLM-based multi-agents (LLM-MAs). Such systems leverage the collective intelligence of specialized agents which act collaboratively by engaging each other in cooperative tasks. LLM-MAs have already been experimented with in various domains, demonstrating their potential in complex problem-solving and world simulation (Guo et al., 2024).

In this vast landscape, LLMs will continue to evolve toward more capable entities at the core of sophisticated AI systems. This fast-paced evolution will see LLMs becoming increasingly pervasive, seamlessly embedded in a wide array of applications, transforming industries and everyday processes by enabling smarter, more adaptive, and efficient solutions.

2.3 Key Principles of Human-Centered Design for LLMs

Human-centered design (HCD) is a problem-solving approach that prioritizes human needs, values, and experiences throughout the design process. The fundamental principle of HCD is to place humans at the center of the design process, engaging them in iterative development and ensuring that technologies align with real-world goals and constraints. Overall, HCD principles offer valuable guidance for developing LLMs that are usable, useful, and aligned with users' goals. According to HCD, the design of human-centered LLMs must begin with a deep understanding of user needs, contexts, and behaviors through comprehensive user research methods such as questionnaires, interviews, and focus groups (Brachman et al., 2024). Iterative prototyping and regular evaluations with a representative set of human subjects are also essential for validating assumptions and refining model behavior and interfaces.

However, HCD principles alone are not sufficient. Because LLMs operate as sociotechnical systems with broad cultural, organizational, and environmental impacts, human-centricity must extend beyond user interaction to encompass systemic and ecological considerations (Ozmen Garibay et al., 2023).

In addition to these general principles, we need other principles that complement and extend HCD in the context of generative AI and LLM. Toward this end, Weisz et al. (2024) proposed six high-level design principles that reflect both reinterpretations of established AI design concerns and novel considerations introduced by generative models. Among these, three principles build on familiar AI challenges but take on new meaning with LLMs:

- *Design Responsibly*: Generative models can produce biased, misleading, or harmful content. Designers must ensure systems are developed and deployed with safeguards for fairness, transparency, and ethical use.
- *Design for Mental Models*: Users often misunderstand how LLMs work, leading to overreliance or misuse. Interfaces should support accurate mental models by clearly communicating the model's capabilities, limitations, and sources of uncertainty.
- *Design for Appropriate Trust and Reliance*: Trust in LLMs must be carefully calibrated. Systems should promote appropriate levels of user confidence by making outputs explainable and flagging uncertainty or low-reliability responses. Three additional principles address the unique characteristics of generative AI:
- *Design for Generative Variability*: LLMs produce varied outputs for the same prompt. Interfaces should allow users to explore alternatives and guide variability productively.
- *Design for Co-Creation*: Generative models enable collaborative content creation. Interfaces should support shared agency between human and machine, allowing users to steer, edit, and co-develop outputs.
- *Design for Imperfection*: LLMs are fallible and may generate flawed or nonsensical content. Rather than hiding imperfections, systems should help users identify, interpret, and manage errors in ways that maintain control and understanding.

3 Challenges

3.1 Human Value Alignment in LLMs

Despite their outstanding potential, LLMs have shown considerable concerns that threaten their trustworthiness, security, and effectiveness. For instance, these models can produce biased content, misinterpret human instructions, generate hallucinations (i.e., factually wrong information), or provide misleading recommendations (Wang et al., 2023b). Therefore, aligning LLMs with human expectations and preferences is being widely explored to establish trustworthy AI-based systems. Human value alignment in LLMs describes the practices of ensuring that LLMs work in a compatible way with human values. Generally, the goal of alignment is adhering to human values deemed important by individuals or society, meaning that AI systems should act in accordance with what is morally projected by people (Shen et al., 2024). As a result, human value alignment is largely being introduced as a main goal to address trustworthiness issues related to LLMs. However, "human value" is a subjective concept that differs according to cultural and societal considerations. Liu et al. (Liu et al., 2023b) have reviewed "human value" in their proposed LLM alignment taxonomy, which consists of seven main categories: reliability, safety, fairness, resistance to misuse, interpretability, social norms, and robustness.

Each of these is divided into more specific topics, resulting in a total of 29 subcategories that represent related issues under each category, such as misinformation under reliability, bias under fairness, and violence under safety. This classification aims to synthesize the various approaches and considerations involved in promoting LLMs' trustworthiness. Shen et al. (Shen et al., 2024) introduced the Bidirectional Human-AI Alignment framework to describe the relationship between humans and AI-based systems regarding human value alignment. This framework has two main directions: (1) Align AI to Humans, which ensures that AI systems reflect human values and intentions, and (2) Align Humans to AI, which focuses on human cognitive and behavioral adaptation toward AI progress and responses. This highlights a collaborative opportunity through which both humans and AI can learn from each other to improve their interactions.

Two main techniques have been proposed to support the generation of more accurate, adequate, and socially acceptable responses: (1) supervised fine-tuning (SFT), which leverages human-provided answers to guide the model to comply with the social norms, and (2) reinforcement learning from human feedback (RLHF), which uses human-labeled feedback to fine-tune the LLM, where labelers rank various responses from the model to identify which are closer to unbiased human answers (Liu et al., 2023a).

In addition to these techniques, system prompts have emerged as an important mechanism for influencing LLM outputs. These are usually inserted prior to the user prompt to steer the model's behavior with the goal of making LLM outputs more safe and helpful (Touvron et al., 2023). Moreover, modern LLMs often incorporate a set of predefined rules or guardrails aimed at preventing harmful outputs (Dong et al., 2024). Although these rules are typically developer-defined and lack transparency, recent research has sought to democratize the process by examining how their design could be informed by broader societal input (Huang et al., 2024).

Despite these efforts, the alignment of LLMs with human values remains an open challenge. One significant issue is the underrepresentation of non-English languages. Most alignment research and techniques rely on English-language data and prompts, raising concerns about the generalizability of these methods to low-resource languages or culturally distinct contexts (Wang et al., 2023b). Additionally, alignment efforts face ongoing difficulties such as modeling diverse and conflicting human values, ensuring long-term consistency in behavior across different tasks and domains, and mitigating unintended model behaviors that may emerge as capabilities scale.

3.2 Addressing Bias and Fairness in LLMs

LLMs can include significant biases that may result in unfair outcomes across various contexts. Identifying and mitigating these biases is crucial for ensuring fair interactions in LLM-based systems, which improve output neutrality and promote informed decision-making. However, due to the complexity of human-AI interaction and the diversity of its associated social implications, addressing bias and fairness concerns in LLMs is considerably challenging (Anthis et al., 2024). Besides, there are many influencing factors standing in the way of promoting LLM fairness, including the lack of determined, universally applicable fairness metrics (Chu et al., 2024; Gallegos et al., 2024; Li et al., 2024c) and the necessity for context-specific solutions to overcome the difficulties in applying traditional fairness frameworks broadly (Anthis et al., 2024). Moreover, detecting biases in LLMs demands expertise across various fields, including software testing, prompt engineering, ethics, and specialized LLM knowledge, to ensure comprehensive and fair bias testing (Morales et al., 2024).

Addressing bias in LLMs requires an understanding of the various sources of bias.

Researchers have identified three main reasons for the presence of bias in LLMs:

- **Data Bias**: Unfairness in data used to pre-train these models arises from unbalanced pre-training datasets that lack equality and diversity in both domains (fields of knowledge) and genres (types of text). Additionally, the selected datasets for pre-training may differ and be influenced by several factors, such as the time of their creation, the people who created them, and the language from which they are derived, which often reflect a specific culture (Kasneci et al., 2023; Navigli et al., 2023).
- **Model Development and Algorithmic Bias**: LLMs can reflect societal biases inherent in the human-generated data on which they have been pre-trained. They may also exhibit patterns resembling those of cognitive bias in humans (Echterhoff et al., 2024). Biases can sometimes be introduced during the LLM development process. Regardless of the model size or complexity level, research shows notable variability in bias scores within different models, depending on their construction. This highlights the critical role of model architecture and training methods in bias mitigation (Kumar et al., 2024a).
- **Social Bias**: This refers to the presence of stereotypes, partialities, and discriminatory attitudes toward specific groups based on different attributes such as ethnicity, religion, physical appearance, and other characteristics. These biases can manifest in various forms, including racism, ageism, and sexism. They often emerge from the historical marginalization of certain groups, resulting in their underrepresentation in training data. Thus, using such a dataset, LLMs may perpetuate these societal inequalities (Navigli et al., 2023).

In the context of LLMs, common bias mitigation techniques used in other AI models, such as prejudice removal, subgroup modeling, or data resampling, are not easily applicable. Instead, a multifaceted approach is required. Mitigation practices can be applied within different stages; accordingly, they are classified into pre-processing, in-training, intra-processing, and post-processing techniques (Gallegos et al., 2024). Ongoing monitoring and testing are also essential to ensure fairness in model outputs through early bias detection and correction (Kasneci et al., 2023).

3.3 Ensuring Transparency and Explainability

As LLMs become increasingly integrated into various applications, ensuring that their decision-making processes are understandable and interpretable is crucial for building trust and accountability. Nowadays, the need for transparency is further emphasized by legal frameworks such as the General Data Protection Regulation (GDPR) in the European Union, which grants individuals the right to an explanation when automated decisions are made by AI-powered systems that can potentially impact them.

To address these challenges, explainable artificial intelligence (XAI) has emerged as a research field focused on providing the why to a certain output produced by complex AI systems, including LLMs. By emphasizing a variety of concepts such as interpretability, transparency, and accountability, the XAI approach seeks to build more trustworthy and human-understandable AI systems from an algorithmic perspective. The black-box nature of LLMs presents a significant challenge to understanding and interpreting their outputs, primarily due to the complexity of their architecture.

With multiple intricate layers that capture various levels of abstraction, it becomes difficult to trace how a specific output is derived. This complexity necessitates the development of explainability methods that can shed light on the internal workings of these models and may provide users with a clear mental model during interaction. Zhao et al. provided an interesting categorization of explainability methods applied to LLMs, specifically in the context of adapting to downstream tasks (Zhao et al., 2024a). Despite ongoing advancements in this field, a significant challenge remains in ensuring that LLMs are transparent and interpretable in a way that is truly meaningful to users. In fact, current explainability approaches are characterized by an algorithm-centric focus, often reflecting the bias of what constitutes an explanation to those who develop and work with these modes, rather than to the end-user (Liao et al., 2020). Therefore, while the algorithm-centric view is still meaningful as it sheds light on the inner workings of complex models, it is not sufficient for fostering genuine understanding or trust among users, particularly those without a technical background. This lack of transparency can potentially lead to confusion, reduced trust, and even rejection of the AI system, especially when users face high-stakes decisions in fields like healthcare, law, or finance. Recent "reasoning" or chain-of-thought models (e.g., ChatGPT o1, Deepseek R1) attempt to mitigate this by generating explicit step-by-step rationales alongside their outputs, thereby attempting to provide users with a transparent view of how the model arrived at its conclusion. However, it is not clear whether these reasoning chains faithfully represent the models' underlying reasoning processes, which calls for further research into the reliability of such explanations (Chen et al., 2025).

3.4 Privacy and Data Protection in LLMs

Ensuring privacy and data security is a key challenge in the development and deployment of LLMs. These models require extensive datasets for training, which may include sensitive or personal data, or data from unverified sources. As a result, LLMs can inadvertently expose sensitive information or undermine the reliability of outcomes by embedding errors, distortions, or misrepresentations. This raises ongoing concerns about data protection, confidentiality, and ethical compliance. Addressing these issues is crucial to building trustworthy and responsible AI systems.

One of the fundamental challenges with LLM privacy is the potential leakage of sensitive data at different stages of development and during use. LLMs may unintentionally expose private information in their outputs. This risk threatens any content that would cause privacy violations if disclosed, including confidential data and personal information such as usernames and email addresses (Das et al., 2025; Li et al., 2024b; Yan et al., 2024a). Sensitive data may sometimes be exposed through user prompts. Users' queries to LLMs may contain personally identifiable information. For example, queries related to medical conditions, personal issues, or financial matters can divulge private aspects of an individual's life (Li et al., 2024b). On the other hand, unlike privacy leakage, where the model's vulnerability enables attackers to access leaked private data, attackers use privacy attacks to breach LLMs to access private data or manipulate model outcomes and control its responses.

Privacy attacks take many forms, including data poisoning attacks, membership inference attacks (Das et al., 2025), attribute inference attacks, model stealing attacks, and model inversion attacks, also known as data reconstruction (Yan et al., 2024a).

3.5 Data Governance in LLMs

When developing LLMs, data privacy and security considerations must be aligned with current legal frameworks, such as GDPR and the California Consumer Privacy Act (CCPA) which both define the frameworks and principles that endorse personal privacy practices and data security (Cilloni et al., 2024). To increase user trust in LLM-based applications, developers can adopt practices that comply with these legislations (Kibriya et al., 2024).

In addition to earlier regulatory efforts, the European Artificial Intelligence Act (EU AI Act)—which entered into force in August 2024—regulates AI systems used within the EU, regardless of where they are developed. It establishes a framework for ethical AI development that applies to both providers and deployers. By focusing on transparency and accountability, the Act is intended to protect fundamental rights, sustainability, and democracy while also promoting innovation and enabling responsible AI deployment (Butt, 2024; Kaffee et al., 2025). The EU AI Act's embedded obligations have already influenced the deployment of LLMs within the region. Notably, Meta opted not to release its multimodal LLaMA model in the EU due to compliance challenges imposed by the Act (Daly, 2025). Similarly, California's SB 1047 specifically regulates the development and fine-tuning of large frontier models, emphasizing developer responsibility to prevent "critical harms." The bill introduces safety measures such as a "kill switch" that allows humans to intervene and prevent models from causing damage, along with independent audits and penalties for noncompliance, with a focus on significant risks such as mass casualties and infrastructure damage (Samuelson, 2025).

Eventually, it is expected that new security threats will emerge over time, which necessitates the continuous development of defense techniques and the regular review of data governance objectives. Enforcing robust regulations against data breaches and security issues is essential for maintaining reliability and ensuring compliance with ethical standards. By adhering to such frameworks, developers can not only enhance transparency and safety in LLM applications but also promote appropriate user confidence toward them.

3.6 Interaction and Accessibility

As previously highlighted, the dynamic capabilities of LLMs make them highly flexible interactive tools for a wide range of tasks. This high degree of versatility is especially enabled by their ability to process user instructions formulated in natural language known as prompts and execute them by generating relevant responses, performing specific actions, or solving problems in alignment with the user's intentions. In this context, the structure and clarity of a prompt are crucial to ensure the correct interpretation of such intentions by the LLM, so it can produce results that align with the desired outcome, maximizing the effectiveness of the interaction.

This shift toward natural language as the primary medium of communication introduces a new interaction paradigm which moves away from traditional interfaces, characterized by elements such as buttons or graphical menus, and instead relies on the conversational dynamics between humans and machines. Through prompting, users can communicate their needs more intuitively and flexibly and with the same degree of freedom as though they were engaged in a natural, human-to-human conversation. However, this paradigm also introduces several interaction challenges. Without clear guidelines or constraints, users may struggle to engage effectively, face misunderstandings, or feel uncertain about how to phrase their requests to achieve the desired outcome. Prompt design challenges, in particular, have emerged as a common difficulty when users attempt to elicit specific behaviors from LLMs. More specifically, Zamfirescu-Pereira et al. (2023) identified several critical user behaviors that shape these challenges. In particular, they noted how a significant number of users initially struggled to understand what types of behaviors could reasonably be expected from the models. This reflects the previously mentioned explainability challenges, where users are unsure of the model's capabilities and how it interprets their instructions. Such uncertainty often leads to frustration, as users may either overestimate or underestimate what the LLM can do, making it difficult to craft effective prompts. In fact, users are unsure of how to iteratively refine prompts to achieve desired results or how to adapt successful prompts for use across varying contexts.

Recent research explores the challenges and dynamics in human-LLM interaction. Subramonyam et al. (2024), for instance, focus on mental models, which can be defined as the internal representations users form about the interactive system.

Mental models guide users in simulating the outcomes of the system's actions, anticipating its behavior, and making decisions about how to engage. In particular, the authors introduce the *gulf of envisioning*, a concept that characterizes the cognitive distance between users' initial intentions and their ability to craft prompts that effectively utilize an LLM's capabilities to generate high-quality output. The gulf of envisioning encapsulates three core challenges: understanding the difference between what the user wishes the LLM to accomplish and what it is inherently able to do (the capability gap), articulating intentions clearly through prompts (the instruction gap), and predicting the LLM's output (the intentionality gap).

While much research on LLMs has focused on their capabilities and user interaction dynamics, less attention has been devoted to making these tools accessible to a wide variety of people, despite **ensuring accessibility and inclusivity** being considered essential in the development of human-centered AI systems (Ozmen Garibay et al., 2023).

On one hand, there is agreement on the great potential of LLMs to be used to improve the lives of people with disabilities. For instance, a 3-month autoethnography suggests that LLM can provide on-demand support for a wide variety of accessibility needs, in low-stakes, easily verifiable contexts (Glazko et al., 2023). Similarly, Valencia et al. (2023) showed that LLMs can produce suggestions that support people who are unable to speak using their voice and rely on AAC devices, potentially saving them time and reducing their physical and cognitive effort when communicating. More recently, Chang et al. (2024) introduced WorldScribe, a system that generates automated live real-world visual descriptions to help blind people understand their surroundings.

On the other hand, in the accessibility domain, most research focuses on LLM-infused tools rather than studying how people with disabilities interact with LLM. In this sense, the current interaction paradigm is far from optimal and presents several challenges. One critical issue is the bias embedded in many LLMs, which can perpetuate harmful stereotypes about disability (Gadiraju et al., 2023). Moreover, current models are seldom tailored to accommodate the unique needs of disabled users. To address these shortcomings, some recent studies are beginning to involve people with impairments directly in the design process. For instance, a recent study engaged blind and visually impaired participants to identify interaction patterns that improve information access via LLM interfaces (Pucci et al., 2024), a participatory approach that represents a promising direction for creating more inclusive and accessible LLM interactions.

4 Designing Human-Centered LLMs

4.1 Human-Centered LLM Design Process

The development of human-centered LLMs requires an integrated approach that gives equal importance to both the design of the underlying model and the design of the user interface (UI) and interaction experience. These two dimensions are interdependent:

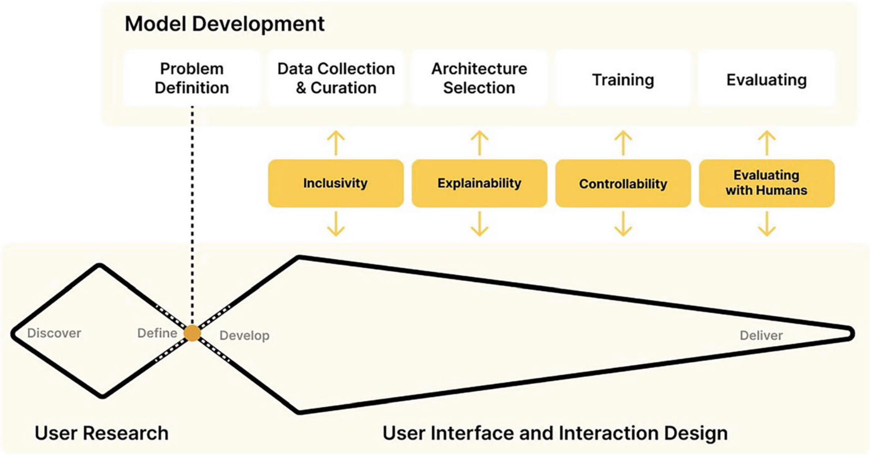

Fig. 1 The integrated approach to the design of human-centered LLMs, combining the language model development workflow with a double-diamond workflow for the UI and interaction design. A shared initial user research phase informs both workflows. Inclusivity, transparency, controllability, and evaluation with human subjects should be considered at every stage of the process

While the model determines the system's capabilities, the interface shapes how users understand, interact with, and benefit from the model. Human-centered LLMs must therefore be designed through a coordinated process that aligns technical functionality with usability, transparency, and inclusivity. This integrated process is illustrated in Fig. 1, which combines the language model development workflow with a double-diamond framework for UI and interaction design. The process begins with a clear definition of the model's intended purpose, target user groups, and application context. User research methods, such as interviews, questionnaires, and focus groups, are employed to identify user needs, expectations, and potential challenges. This early engagement informs both model capabilities and interaction design, ensuring that the system is aligned with real-world use cases and remains inclusive and accessible (Norman, 2013).

In the **model development workflow**, the next phase involves data collection and curation. The quality, diversity, and ethical sourcing of data play a crucial role in shaping model behavior. Data must be representative of various demographic, cultural, and linguistic groups to reduce biases and support generalization across diverse contexts. Documentation of data sources, legal compliance, and transparency are necessary to ensure accountability and traceability (Liao & Wortman Vaughan, 2024).

After data collection and curation, the next phase involves training the LLMs. This phase includes both pre-training—in which the model learns general language patterns from large-scale datasets—and post-training, where task-specific fine-tuning and alignment methods are applied. Post-training may involve supervised fine-tuning, reinforcement learning from human feedback (RLHF), or other alignment techniques aimed at refining the model's responses to be more safe, helpful, and aligned with user expectations. During this phase, attention must also be given to sustainability:

The environmental impact of training large models is considerable, and where possible, smaller or more efficient architectures should be considered (Singh et al., 2024).

To ensure reliability and user alignment, mechanisms that support controllability should be integrated (Liang et al., 2024). Finally, evaluation with human subjects is essential to assess the system's real-world performance. Such evaluations help identify usability issues, fairness concerns, and potential risks, supporting iterative refinement. In parallel, the **UI and interaction design process** follows a similarly human-centered approach. The initial phase of formative studies—discussed earlier—provides valuable insights into the expectations and needs of those who will interact with and be affected by the system. Accessibility considerations must be addressed from the outset to ensure that users with diverse abilities can engage with the interface effectively. Interaction design should prioritize explainability by including features that clarify model outputs. These features help users understand how the system operates and where its limitations lie, fostering trust and supporting responsible use. The UI should also promote user agency and control. It should help individuals form accurate mental models, identify and manage errors, and remain in command of how the system is used. This includes enabling users to guide, adjust, and refine the system's outputs as needed. By making system behavior more transparent, the interface can help establish an appropriate level of trust (Weisz et al., 2024).

After prototyping, the interface is evaluated using both automated methods and human subject evaluation. The insights gained from these evaluations drive continuous enhancements in functionality and overall user experience. Several components span both the model and UI design workflows. For instance, evaluating with human subjects is critical at multiple stages across both layers. Similarly, efforts to promote inclusivity, support controllability, and provide explanations require close coordination between the people who are developing the model and those who design its interface.

In summary, the successful development of human-centered LLMs depends on an integrated design approach in which model architecture, training, and alignment are developed in tandem with the design of how people should interact with the system. Such an approach permits to design human-centered LLMs, systems that are not only capable and efficient but also transparent, inclusive, and aligned with human needs and values.

4.2 Data Collection and Curation

The development of high-performance LLMs is inherently influenced by the quality, scale, and diversity of the datasets utilized across each training phase, from pre-training to fine-tuning and alignment. While scaling intelligence and performance is indeed critical as a direction to follow in order to enhance the emerging capabilities of LLMs, it is equally important to consider the impact of such capabilities on broader societal and ethical concerns. Recent research demonstrates how LLMs trained on vast amounts of data do not automatically acquire a consistent understanding of diverse cultural values (Kharchenko et al., 2024). These findings underscore the importance of an inherently human-centered dataset construction pipeline that proactively incorporates and accurately represents the diversity of human values.

In this light, the phase of data collection should incorporate data that has been thoroughly evaluated for cultural diversity. Given the variegated spectrum of human identities, human values are not to be considered monolithic, as they are shaped by a wide range of cultural, social, and individual factors. An interesting perspective on the categorization of human values emphasizes that human values are influenced not only by individual experiences but also by the broader social and interactive contexts in which humans operate (Shen et al., 2024). Such an evaluation must not only focus on diversity but must also revolve around preventing the assimilation of cultural misunderstandings and stereotypes (Kharchenko et al., 2024). Moreover, to ensure responses are based on reliable factual cultural understandings, LLMs should reference qualified sources also when making cultural assumptions, for example, by applying Retrieval-Augmented Generation (RAG) pipelines specifically integrating a knowledge base of cultural dimensions (Hofstede, 1984).

Prompt engineering techniques have been explored to elicit specific cultural perspectives from LLMs; however, they have been proven to be ineffective. Another line of work involves pre-training and fine-tuning techniques that, however, require extensive datasets and significant computing resources, making it costly and difficult to apply to many cultures, especially underrepresented ones. Alternative solutions propose semantic data augmentation as a cost-effective solution (Li et al., 2024a).

Data curation pipelines should also address concerns about transparency and privacy as part of the LLM development lifecycle. Given the massive scale and diverse sources of data that lack proper documentation, achieving transparency in data curation pipelines is a significant challenge. For this reason, efforts should be made in favor of model reporting strategies that include clear information about the data used for pre-training and fine-tuning phases. Guidelines should focus on distilling and communicating the most critical characteristics of these datasets to provide a basic understanding of what the models are trained on (Liao & Wortman Vaughan, 2024). Furthermore, as companies collect user data destined for model training, it is crucial to safeguard individual and collective privacy. This can be achieved by following a privacy by design approach throughout model development and deployment phases. Moreover, providing transparency to users about how their data is collected, used, and stored is essential for fostering trust. Finally, more proactive human-centered approaches would eventually empower users with control over their data, including the ability to understand its impact on models and to potentially revoke consent for its use (Ozmen Garibay et al., 2023).

4.3 Providing Explanations

To place humans at the center of LLM-based solutions, it is crucial to ensure a meaningful level of transparency that fosters user trust. Although recent advances such as the development of reasoning models (see Sect. 3.3) have begun to address this challenge, a more coordinated and comprehensive effort is still required. In this direction, explainable artificial intelligence (XAI) has made strides to uncover a variety of approaches able to provide a sense of transparency from an algorithmic perspective. While these more traditional approaches are indeed valuable to obtain tangible insights about the internal reasoning of a diverse plethora of models, they fail to prioritize the needs and understanding of the human user.

Human-centered XAI (HCXAI) seeks to carry out such a paradigm shift in this context, hence by moving beyond the singular focus on algorithmic transparency. The HCXAI perspective recognizes that explaining complex systems like LLMs requires the consideration of not only how the "black box" works internally, but also who is interacting with it and why they seek explanations. In this sense, HCXAI seeks to highlight that the individual interpreting the explanation is as important as the explanation itself (Ehsan & Riedl, 2024). For example, recent findings have highlighted the heterogeneity of user needs, as AI experts and nonexperts interpret explanations differently and have distinct requirements for what constitutes a human-like explanation (Ehsan et al., 2024b). Furthermore, the motivations behind seeking explanations from LLMs are diverse, ranging from building trust and understanding causality to evaluating the AI's capabilities (Liao et al., 2020). Consequently, effective XAI for LLMs must move beyond generic explanations and consider the specific user, their background, and their reasons for seeking explanations.

HCXAI's proposal is to empower users by working toward promoting the calibration of user trust, as opposed to its simple enhancement. This strategy would ensure that user trust aligns with the model's actual capabilities and limitations. For example, in the specific context of LLMs, expressing uncertainty has been shown to reduce overreliance, indicating a path toward better trust calibration (Zhang et al., 2020). A strategy that seeks to help users better calibrate their trust, combined with the provision of context-aware explanations that address specific user needs and backgrounds, would better equip individuals in the assessment of the reliability of LLM outputs.

In fact, as LLMs are renowned for their tendency to generate nonfactual responses (i.e., hallucinations), addressing the inherent complexities and occasional fallibility of LLMs is crucial to building effectively calibrated user trust. In this sense, the concept of Seamful XAI emerges as a valuable design approach (Ehsan et al., 2024a). This framework proposes that instead of concealing the limitations or "seams" of AI systems, strategically revealing them can enhance explainability and user agency. By making the imperfections of LLMs visible, users can develop a more realistic understanding of their capabilities and potential failure points. In a more practical sense, implementing Seamful XAI involves the identification of potential breakdowns, the strategic anticipation and incorporation of seams in order to highlight failures, and the design of user interfaces that successfully incorporate these seams.

By combining such emerging strategies with design principles that promote human-centricity in the development and evaluation of novel XAI approaches, researchers and practitioners can create more transparent, human-centered LLM interactions that foster appropriately calibrated trust, ultimately empowering users to engage with AI systems more effectively and responsibly.

4.4 Interaction and Control

As LLMs become increasingly integrated into various human workflows, addressing the inherent challenges necessitates the implementation of a multifaceted strategy encompassing thoughtful design approaches, as well as the creation of intuitive interaction paradigms.

With their interaction paradigm based on natural language, LLMs substantially narrow the gulf of execution typical of conventional software systems. However, users are introduced to the new challenge of formulating their intentions into effective prompts that can leverage the LLM's capabilities to produce the desired output (Subramonyam et al. 2024). In this light, Subramonyam et al. (2024) suggested new design patterns intended to support users in the formation of effective mental models:

- Visually tracking prompts and outputs: Users should be provided with effective visual interfaces that enable them to see their prompts and the resulting outputs to better understand the consequences of different prompt strategies.
- Suggest ideas for prompting: Users who are not AI experts often tend to struggle in obtaining effective prompts (Zamfirescu-Pereira et al., 2023). Proactive systems that offer prompt suggestions would help users unfamiliar with LLMs in prompt engineering tasks or, eventually, serve as inspiration.
- Provide multiple outputs: Instead of generating just one output, offering several different outputs in response to a single prompt would allow users to compare options and see which best aligns with their intentions.
- Use domain-specific prompting strategies: Prompting techniques customized for tasks specific to a certain domain would help users craft more effective prompts in terms of relevancy while reducing ambiguity.
- Allow manual control of the input: This pattern emphasizes giving users the ability to manually edit the outputs and interactions with LLMs. Direct manipulation of the outputs has been proven to effectively help users in the incorporation of their values and intentions into their prompts (Masson et al., 2024).

An important necessity is to actively foster and protect innate human skills and prevent deskilling. In fact, with the introduction of LLMs, the risk of diminished critical engagement and overreliance is becoming more pervasive, making it crucial for users to actively evaluate and engage with generated content. To effectively protect critical thinking, several design implications have been suggested (Lee et al., 2025):

- Enhancing awareness of critical thinking opportunities: Generative AI tools should proactively highlight the need and opportunity for critical thinking, especially in situations where it might be overlooked, such as for tasks perceived as unimportant. Moreover, users could be empowered to explicitly request assistance on critical thinking tasks when they consciously need it.
- Increasing the motivation to think critically: Generative AI could be exploited as a way to contribute to long-term skill development. In fact, tools that help users in

the critical evaluation of outputs can improve the user's own understanding and abilities.

- Enhancing the ability to execute critical thinking: Design efforts should facilitate user learning about how to effectively evaluate and improve generated responses. The provision of explanations can help achieve this by offering insights into the process behind the output and, consequently, help users in the assessment of its validity and subsequent refinement.

4.5 Ensuring Accessibility and Inclusivity

The widespread adoption of LLMs has made it imperative to ensure that future LLM-based applications are accessible and inclusive, benefiting all users regardless of their abilities, linguistic background, or cultural context. If such factors are not addressed from the initial stages of development, LLM-based applications risk inadvertently exacerbating existing biases and digital divides that hinder equitable access to information.

In the text context of inclusivity, an LLM should effectively cater to the diverse needs and preferences of all users, considering linguistic and cultural diversity. As an example, LLMs have demonstrated varying degrees of multilingual proficiency, with performance disparities in favor of the English language. This performance gap can significantly impact the user experience for individuals whose primary language is not English in terms of performance and bias (Hoadley & Vogel, 2024; Biswas et al., 2025). In this sense, an inclusive LLM should support multiple languages, understand and respect different cultural contexts, and avoid perpetuating biases against any group.

Data diversity has a foundational role in mitigating biases within LLMs. By including representations from various demographic groups, including individuals with disabilities, different genders, ethnicities, and socio-economic backgrounds, it is possible to reduce the likelihood of the model perpetuating harmful biases (Li et al., 2024a). However, while data diversity is a vital step, it is important to recognize that it may not be sufficient on its own. Future research should focus on fostering interdisciplinary collaboration between NLP and HCI researchers, with the goal of developing novel approaches for aiding in the active identification and mitigation of biases. For example, interfaces could be developed in a way that ensures that biases are not only detected and brought to the user's attention, but also actively mitigated throughout interactions.

In the context of designing accessible LLMs, the design of these models and their interfaces must be tackled in a way that allows for effective use by individuals with a diverse range of abilities, encompassing visual, auditory, motor, and cognitive abilities. This includes ensuring compatibility with assistive technologies such as screen readers, providing alternative input and output methods, and designing interactions that are intuitive and understandable for users with varying cognitive abilities.

Considering the possibilities of information access with LLMs, existing conversational patterns for web navigation can be potentially transposed to prompt-based interactions with LLMs. For example, an accessible prompt-based interface should allow users to seamlessly understand the topic of the chat, enable navigation through conversation history, provide overviews, allow bookmarking of specific content, and organize information into segments and topics for easy exploration and retrieval (Pucci et al., 2024).

Additionally, the development of multimodal LLMs holds significant promise for enhancing accessibility by supporting diverse communication needs beyond just text, such as by providing seamless voice interaction for visually impaired individuals or by allowing the understanding and generation of sign language.

4.6 Evaluating with Human Subjects

When approaching the evaluation of human-centered LLMs, it is important to recognize that these reflect complex goals that cannot be properly reached or assessed without the active involvement of human subjects (Ozmen Garibay et al., 2023). While there is a long tradition of NLP techniques to evaluate generative models (Sai et al., 2022), traditional evaluation metrics, which often prioritize accuracy and utility, are proving insufficient to fully capture the complexities of human interaction with LLMs. In such cases, standard metrics often fall short, requiring more complex measures and imperfect "ground truth" references, which are frequently chosen for convenience or created by crowd workers. An additional challenge arises from the inherently conversational nature of LLM interactions, as a model's performance may depend not on a single input-output pair, but on the progression of several turns in a dialogue. This makes evaluation even more complex, as the quality of an output may only become evident after multiple interactions.

To address these challenges, direct involvement of human subjects is essential and should complement automated methods. The evaluation could comprise tasks where people assess outputs based on criteria like quality, fluency, coherence, relevance, adequacy, or informativeness (Liao & Wortman Vaughan, 2024).

Despite these needs, one of the most commonly used paradigms remains *LLM-as-a-judge*, where large language models are used to evaluate the outputs of other LLMs (Gu et al., 2024). While this approach offers scalability, it is limited by the lack of direct human involvement. To address this, recent approaches are increasingly incorporating human input into the evaluation process.

One notable example is EvalLM, a system that supports the evaluation of outputs on user-defined and application-specific criteria (Kim et al., 2024). EvalLM uses an LLM in two roles: (1) as an evaluation assistant that scores outputs based on user-defined criteria and (2) as a criteria reviewer that helps improve those criteria. The evaluation assistant explains its judgments to help users understand where outputs fell short or where the criteria might have been misunderstood. The criteria reviewer suggests ways to refine the evaluation criteria, making them more detailed and accurate. By repeating this process, EvalLM supports users in gradually improving both their prompts and evaluation criteria, resulting in better-quality applications.

Some of the major attempts to directly involve humans in LLM evaluation efforts are exemplified by platforms such as Chatbot Arena, which benchmarks LLMs through anonymous, randomized pairwise comparisons, where users vote on preferred outputs (Chiang et al., 2024). At the time of writing, this crowdsourced approach has accumulated over 2.9 million comparisons, producing a dynamic leaderboard that reflects real-world user preferences.

Despite the many research efforts in devising effective evaluation methods, current LLM evaluation practices still face a critical gap due to the absence of standardized, human-centered frameworks. Bridging this gap calls for community engagement in both evaluation and auditing processes. As a result, we are currently witnessing new initiatives that aim to address this challenge by bringing together HCI and AI researchers and practitioners to explore inclusive, stakeholder-driven methods that better reflect real-world needs, uncover potential risks, and align with the values of impacted communities (Xiao et al., 2024).

5 Case Studies of Effective User Interaction with LLMs

5.1 DirectGPT: Enhancing LLM Interaction Through Direct Manipulation

This case study examines the development and evaluation of DirectGPT (Masson et al., 2024), a novel interface designed to improve human interaction with LLMs by integrating principles of direct manipulation.

Traditional LLM interfaces rely on text prompts, which need to be iteratively refined often by making precise references to elements in the output. DirectGPT allows users to interact directly with the LLM's output through physical actions and visual feedback with the goal of making interactions more intuitive, efficient, and controllable. To achieve this goal, DirectGPT seeks to apply the principles of direct manipulation (Shneiderman, 1984). The continuous representation of the object of interest is implemented by ensuring that generated objects (text, code, images) are always visible and accessible. DirectGPT allows users to interact through actions like selecting, dragging, and clicking, alongside or instead of solely relying on text prompts, thus enabling users to tackle their tasks by means of physical actions or labelled button presses instead of using complex syntax. Finally, DirectGPT enables users to perform rapid, incremental, reversible operations whose impact on the object is immediately visible by incorporating undo/redo mechanisms for reversibility and by providing immediate feedback on the targeted elements.

By integrating such principles, DirectGPT seeks to tackle some key issues of traditional prompting. In fact, direct manipulation helps in reducing ambiguity by enabling users to directly refer to specific objects; thus, users can better constrain the scope of their prompts. Moreover, by integrating undo and redo mechanisms, users gain greater control over their interactions, allowing them to iterate through their actions and recover previous states. DirectGPT also introduces a toolbar of reusable, templated prompts, making the repeated application of actions to different objects more rapid and efficient.

DirectGPT works by transforming direct manipulation actions into engineered prompts that are then sent to the underlying LLM. It uses strategies such as sending only the selected part for rewriting when localizing effects and using delimiters or unique IDs to refer to objects dragged into the prompt.

This approach led to higher user satisfaction and a perceived increase in control compared to a purely conversational interface. In a user study, DirectGPT was proven to help users be significantly faster, more efficient, and more successful in their tasks compared to a baseline ChatGPT interface. In fact, participants using DirectGPT completed tasks by using significantly shorter prompts and with fewer interactions; hence, they were significantly faster while also achieving higher success rates.

5.2 AI Chains: Transparent Human-AI Interaction with Prompt Chaining

Another effective case study is the implementation of AI Chains (Wu et al., 2022). When tackling complex multistep tasks, users often need to ensure the LLM interprets and performs the task accurately. For this reason, crafting an effective prompt is crucial to ensure that the model correctly captures all the nuances required to complete the task successfully.

However, while crafting natural language prompts may seem intuitive, it becomes much more challenging when the task involves multiple steps or requires specific instructions. For users who lack a strong mental model, it may be cognitively demanding to write and tweak effective prompts. This issue is also exacerbated by the fact that LLMs have demonstrated limited accuracy when performing tasks that require a sustained reasoning process through which they can gradually accumulate imperfections (i.e., exposure bias), resulting in a gradual decline of performance as the context lengthens.

By drawing parallels with the workflows of crowdsourcing that use the concept of "microtasking," this work proposes a prompt chaining approach by which the complex task is broken down into smaller sub-steps subsequently described by more manageable prompts which are then logically linked together. Moreover, they couple this method with a user interface purposefully designed to give the users a clear overview of the prompt chain structure and to allow for the seamless tracking, modification, and refinement of each step involved in the process.

This modular approach has been proven to enhance the key challenges previously mentioned in this chapter, with a specific focus on controllability, system transparency, and debugging. The structured nature of the approach allows users to target specific sub-tasks independently; for this reason, it reduces cognitive load by enabling a more granular approach to prompting. This subsequently leads users to construct a more accurate mental model of the capabilities and limitations of the system. Finally, this approach enhances system transparency as it provides users with a clear and increased visibility of how the system operates; hence, it provides a more interpretable overview of the reasoning process.

6 Emerging Trends and Future Directions

6.1 The Future of LLMs in Scaling Intelligence and Performance

The continuous increase in the number of trainable parameters and the vastness of the datasets used for training have so far been fundamental drivers behind the advancements observed in LLM performance and the emergence of new capabilities. This expansion in size directly correlates with increasingly sophisticated capabilities, such as reasoning and more effective alignment for understanding human intent in a variety of tasks. However, this trend of increasing scale does not come without limitations and new challenges to consider.

Scaling the number of hardware accelerators for training quickly encounters diminishing returns, as the overhead from distributed communication strategies becomes significant. These factors suggest that alternative approaches may be necessary for continued advancement (Fernandez et al., 2024). Moreover, pre-training large language models entails significant environmental costs concerning energy consumption, subsequent carbon emissions, water usage, and other environmental impacts (Morrison et al., 2025). At the same time, inference also plays a non-negligible role in the overall computational costs (Samsi et al., 2023). Moreover, approaches that employ multistep reasoning can significantly increase per-query computation costs, given their need for generating additional tokens dedicated to the reasoning process.

In this light, the use of smaller and more efficient LMs is gaining traction for domain-specific applications, and techniques like knowledge distillation and layer pruning are being investigated to create more efficient models (Gromov et al., 2024).

The automation capabilities of LLMs naturally raise questions about which human tasks and roles might be significantly altered. Generative AI tools, including LLMs, are currently influencing the work and cognitive processes of knowledge workers. In this sense, human effort is moving from task execution to task stewardship (Lee et al., 2025). This proves how the nature of the human role is evolving, requiring a shift in skills and a careful consideration of how humans and AI can best work together to maintain quality and avoid overreliance.

Looking ahead, the opportunities for progress are immense in terms of the augmentation of human capabilities, driving innovation across various sectors. However, these opportunities must be pursued with a careful awareness of the potential pitfalls. Continued focus on human-centered design principles, sustainable practices, and robust mechanisms for value alignment will be essential to promote a future where LLMs are both beneficial and ethically sound.

6.2 Toward Mutual Alignment in Human-AI Collaboration

One major perspective in HCAI is that ensuring AI systems are valuable and aligned with human values and requirements. As such, integrating human-centered principles is critical for refining AI systems' usability and trustworthiness (Shneiderman, 2020). Setting the foundations for how humans and AI can work together is crucial, especially as AI technology continues to evolve (Sun et al., 2024). To maximize the benefits of aligning AI with users' expectations, adopting an ongoing mutual alignment approach is recommended. As briefly mentioned in Sect. 3.1, Shen et al. (2024) introduced the Bidirectional Human-AI Alignment framework, which aims to adopt a collaborative approach by establishing mutual understanding and communication between humans and AI systems, allowing for more effective interactions from both sides. This framework consists of two main directions:

1. Aligning AI to Humans: This aims to ensure that AI systems are designed to meet human requirements and to reflect their values by integrating human specifications into the different phases of the AI development process. This involves understanding how to identify and specify human values for AI alignment and how these values can be effectively adopted by AI systems.
2. Aligning Humans to AI: When people interact with AI systems, these systems influence their thoughts and behaviors. To manage this impact and interact effectively and ethically, it is necessary to understand and adapt to AI advancements. This involves helping humans adjust their comprehension and modulate their behavior in response to these developments. The strategy includes educating people about AI, encouraging critical thinking about its outputs, and ensuring effective interaction by assessing the reliability and ethical implications of AI outputs.

In a similar context, Pyae (2025) proposed the "Human-AI Handshake Model," a framework that aims to form a dynamic partnership that enables AI to be a responsive partner instead of being limited to a supportive role and encourages its development alongside users over time. This bidirectional framework includes five essential characteristics: (1) information sharing: for adaptive, two-way communication; (2) mutual learning: for continuous, interactive process; (3) mutual capability augmentation: for a synergistic partnership; (4) feedback: for a collaborative learning; and (5) validation: for a mutual verification and adaptation. This approach can establish more dynamic collaboration and promote more balanced interaction. However, some challenges must be handled for more effective cooperation, such as the need for ongoing user oversight, limited learning abilities, and ethical implications.

Eventually, adopting mutual alignment strategies in human-AI collaboration can lead to more effective, transparent, and reliable interactions. Enabling LLMs to adapt to user preferences and expectations supports human decision-making and enriches bidirectional collaboration. Simultaneously, helping users understand AI behaviors can foster trust and promote more responsible interactions.

6.3 Prospects of Human-Centered LLMs in Various Sectors

LLMs are impacting diverse domains such as education, healthcare, knowledge work, and business, where they can support more effective communication, decision-making, and automation.

In education, LLMs can offer personalized learning experiences tailored to individual needs, support teaching activities, facilitate assessment and grading, and provide real-time feedback. Given their potential to generate human-like text and perform a variety of cognitive tasks, these models can be helpful in creating and refining educational materials, as well as enhancing student engagement through interactive learning processes (Chan & Hu, 2023; Kasneci et al., 2023; Yan et al., 2024b). However, the integration of LLMs into educational practices presents significant challenges, including the potential for both overt and implicit biases in model outputs (Warr et al., 2024), the risk of misuse, the ongoing necessity for human oversight, low levels of technological readiness in certain contexts, and ethical concerns such as transparency deficits and privacy issues (Kasneci et al., 2023; Yan et al., 2024b). To mitigate these challenges, it is essential to adopt responsible AI practices within educational settings with an emphasis on raising awareness about both the risks and potential benefits of LLMs among stakeholders —particularly educators and students—who must develop an understanding of the models' limitations as well as their capabilities. This requires fostering critical thinking skills and applying effective strategies for knowledge acquisition and evaluation (Kasneci et al., 2023).

In healthcare, one of the earliest and most widely adopted applications of LLMs regards the automation of clinical documentation, where AI scribes transcribe and summarize patient encounters to reduce administrative workload and support clinician efficiency (Tierney et al., 2025). More broadly, LLMs have introduced unprecedented contributions across various aspects of care, including pre-diagnosis, diagnosis, patient follow-up, and medical decision-making (Mondal & Mondal, 2024), for instance, by answering patient questions at a level very similar to that of healthcare providers (Nov et al., 2023).

Multimodal LLMs (M-LLMs) could make complex medical analyses easier to understand and interpret by converting complicated visual or auditory data into text-based concepts, which facilitate the interactions among physicians and practitioners with different types of content (Dang & Jia, 2025; Meskó, 2023). LLMs can also support telehealth, from scheduling appointments to teleconsultations, incorporating clinical decision support, precisely interpreting symptoms, and providing personalized medical advice (Kwan, 2024). However, research communities are still trying to address the technological, legal, and ethical implications when integrating these models into the healthcare sector, as their irresponsible use could lead to catastrophic consequences given the high sensitivity of this field. The potential for errors in the outputs underscores the need for ongoing supervision by human experts in the medical domains. Furthermore, concerns such as safeguarding data privacy and ensuring the trustworthiness of AI-generated recommendations are still important challenges that call for more research (Reddy, 2024).

The adoption of LLMs in business fields offers significant advancements in process efficiency and personalization. In finance, for instance, LLMs such as

GPT-4 are used to automate tasks like predicting market trends, analyzing investors' perspectives on the market, generating financial reports, and providing users with personalized financial guidance. These models can process and summarize large amounts of financial data, helping companies make better investment decisions and improve their operations and customer service levels (Pathak et al., 2025; Yang et al., 2024; Zhao et al., 2024b). A key concern with LLMs in finance is their potential to generate hallucinations or inaccurate outputs. Moreover, the cost and complexity of consistently updating these models to remain aligned with changing financial regulations and unstable market and economic situations present considerable challenges. As such, integrating accurate and up-to-date knowledge into LLMs is essential for fostering the reliability of financial decision-making (Yang et al., 2024). This stresses the need for a phased implementation approach, which involves stakeholder engagement, thorough evaluation, and ongoing learning processes (Xu, 2024).

In customer service, LLMs have made promising contributions by enabling adaptive chatbots and virtual assistants to manage a wide array of inquiries, offer personalized advice, and provide real-time support. This reduces the burden on human agents, cuts some of the operational costs, and, most importantly, improves customer satisfaction. Additionally, LLMs like ChatGPT and Claude 3 can help organizations work more efficiently, especially through analyzing customer feelings (sentiment analysis) and extracting useful information from user-generated content (Falatouri et al., 2024). However, challenges remain beyond privacy and data protection concerns, as well as potential hallucinations, such as contradictions between human opinions and predictions made by the models. This highlights the continued need for human oversight to ensure reliable results (Falatouri et al., 2024). One way to overcome these challenges is by leveraging pre-trained models, integrating customer feedback, and encouraging ongoing education for workers (Xiaoliang et al., 2024).

In the field of knowledge work, LLMs are changing the way people handle everyday digital tasks. Previous studies have shown that using different contextual signals can improve search functions for both personal (Jacucci et al., 2021) and external information (Vuong et al., 2021a, b), as well as enhance collaboration (Andolina et al., 2018b). Moreover, recommendations and visual expression of intent can improve effectiveness and efficiency in information-intensive tasks, (Andolina et al., 2015, 2018a; Klouche et al., 2015), potentially freeing more time for reflection (Andolina et al., 2018a). With the advent of LLMs, leveraging context is becoming easier. In particular, the Retrieval-Augmented Generation (RAG) technique does this by fetching relevant internal documents at query time and feeding them into the prompts to produce more accurate and grounded answers (Veturi et al., 2024). This is making internal enterprise knowledge more accessible to both employees and customers, opening up a wide range of opportunities, but also challenges, such as the erosion of critical thinking and perceived agency (Alrabie et al., 2025; Butler et al., 2024). In response to these challenges, recent work is increasingly drawing on earlier research on recommendations, direct manipulation, and visual expression of intent to guide the design of more intuitive and reflective interactions with LLM-powered systems that assist knowledge workers in information-finding tasks and beyond (Masson et al., 2024; Mellatdoust et al., 2025; Pardo Gutierrez et al., 2025).

7 Conclusions

7.1 Summary and Key Insights

This introductory chapter provides an overview of human-centered LLMs, addressing current challenges and outlining foundational design principles from recent research. The following key insights emerge:

- LLMs offer substantial potential to empower users across various domains. However, achieving their full benefits requires them to evolve into human-centered LLMs, systems that seek to guarantee safety, alignment with human values, and ethical standards.
- Effective human-centered LLM design necessitates an integrated approach that focuses not just on the model itself but also on carefully crafted interfaces and interaction methods.
- Human-centric design principles should inform every stage of the model's life cycle—from initial development to deployment, continuous evaluation, and refinement.
- Inclusivity and accessibility are critical in designing human-centered LLMs and related applications. Linguistic and cultural diversity should be considered throughout both model development and interface design. Additionally, user interfaces must ensure compatibility with assistive technologies to fully accommodate users with diverse abilities and needs.
- Traditional automated evaluation metrics are inadequate for comprehensively capturing the nuances and complexities of human-LLM interactions. Therefore, evaluating with human subjects is vital throughout the entire development process.
- Integrating robust mechanisms for controllability and explainability is crucial. This ensures that users understand model behaviors, can easily manage or correct outputs, and maintain effective control over interactions, thus fostering transparency, trust, and user empowerment.
- The use of larger models should be limited to cases where their added capabilities offer justifiable benefits, while smaller models should be preferred in most situations due to their adequacy for many tasks and better alignment with sustainability considerations.

7.2 Final Thoughts

Designing human-centered LLMs is a challenging journey, but it offers the opportunity to design a future where technology better serves humans and the world they live in. To reach this goal, we are encouraged to reimagine the boundaries of AI, ensuring that as new models advance, they remain anchored in principles that prioritize human dignity, collective progress, and sustainable practices. This approach requires collaboration among researchers, designers, and communities to address real-world challenges and promote a future in which LLMs and their related applications can contribute positively to society. Ultimately, by keeping human values at the core of LLM development, we can guide this technology in a way that enhances quality of life and promotes an inclusive and sustainable future.

Competing Interest Declaration The author(s) has no competing interests to declare that are relevant to the content of this manuscript.

References

Alrabie, L., & Andolina, S. (2025). Towards Human-Centered RAG: A study of AI-supported testing practices in italian public administration. In *16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), Salerno, Italy*. ACM, New York, USA. https://doi.org/10.1145/3750069.3750103

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–13).

Andolina, S., & Konstan, J. A. (2023). *Introduction to the special issue on AI, decision-making, and the impact on humans*. Taylor & Francis.

Andolina, S., Klouche, K., Peltonen, J., Hoque, M., Ruotsalo, T., Cabral, D., et al. (2015). Intentstreams: smart parallel search streams for branching exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 300–305).

Andolina, S., Klouche, K., Ruotsalo, T., Floréen, P., & Jacucci, G. (2018a). Querytogether: Enabling entity-centric exploration in multi-device collaborative search. *Information Processing & Management., 54*(6), 1182–1202.

Andolina, S., Orso, V., Schneider, H., Klouche, K., Ruotsalo, T., Gamberini, L., et al. (2018b). SearchBot: Supporting voice conversations with proactive search. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 9–12).

Anthis, J., Lum, K., Ekstrand, M., Feller, A., D'Amour, A., & Tan, C. (2024). *The impossibility of fair LLMs*. https://arxiv.org/abs/2406.03198

Bellogín, A., Grau, O., Larsson, S., Schimpf, G., Sengupta, B., & Solmaz, G. (2024). The EU AI act and the wager on trustworthy AI. *Communications of the ACM, 67*(12), 58–65.

Bilgram, V., & Laarmann, F. (2023). Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods. *IEEE Engineering Management Review, 51*(2), 18–25.

Biswas, S., Erlei, A., & Gadiraju, U. (2025). Mind the gap! Choice independence in using multilingual LLMs for persuasive co-writing tasks in different languages. arXiv preprint arXiv:250209532.

Brachman, M., El-Ashry, A., Dugan, C., & Geyer, W. (2024). How knowledge workers use and want to use LLMs in an enterprise context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–8).

Butler, J., Vorvoreanu, M., Janssen, R., Sellen, A., Immorlica, N., Troy, A., et al. (2024). *Microsoft new future of work report 2024*. Microsoft.

Butt, J. (2024). Analytical study of the world's first EU artificial intelligence (AI) act. *International Journal of Research Publication and Reviews., 5*(3), 7343–7364.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education., 20*(1), 43.

Chang, R. C., Liu, Y., & Guo, A. (2024). WorldScribe: Towards context-aware live visual descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18).

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman. J., et al. (2025). Reasoning models don't always say what they think. arXiv preprint arXiv:250505410.

Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., et al. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first international conference on machine learning*.

Chu, Z., Wang, Z., Zhang, W. (2024). *Fairness in large language models: A taxonomic survey.* https://arxiv.org/abs/2404.01349

Cilloni, T., Fleming, C., & Walter, C. (2024). You are what you buy: Personal information extraction from anonymized data. *IEEE Access, 12*, 29714–29722.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., et al. (2023). The future landscape of large language models in medicine. *Communications Medicine, 3*(1), 141.

Dalsaniya, A., & Patel, K. (2022). Enhancing process automation with AI: The role of intelligent automation in business efficiency. *International Journal of Science and Research Archive, 5*(2), 322–337.

Daly, K. (2025). *The geopolitics of AI regulation*. https://yris.yira.org/global-issue/the-geopolitics-of-ai-regulation/. Yale Review of International Studies (Online), Global Issue, Apr 9 2025.

Dang, T., & Jia, H. (2025). Multimodal large language models in human-centered health: Practical insights. *IEEE Pervasive Computing, 23*(4), 87–93.

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys, 57*(6), 1–39.

Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition C&C'22* (pp. 623–627). Association for Computing Machinery.

Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., et al. (2024). Building guardrails for large language models. arXiv preprint arXiv:240201822.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. arXiv preprint arXiv:240300811.

Ehsan, U., & Riedl, M. (2024). Explainable AI reloaded: Challenging the XAI status quo in the era of large language models. In *Proceedings of the Halfway to the Future Symposium* (pp. 1–8).

Ehsan, U., Liao, Q. V., Passi, S., Riedl, M. O., & Daumé, H. (2024a). Seamful XAI: Operationalizing seamful design in explainable AI. In *Proceedings of the ACM on Human-Computer Interaction*. 8(CSCW1). https://doi.org/10.1145/3637396.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I. H., Muller, M., et al. (2024b). The who in XAI: How AI background shapes perceptions of AI explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–32).

Falatouri, T., Hrušecká, D., & Fischer, T. (2024). Harnessing the power of LLMs for service quality assessment from user-generated content. *IEEE Access, 12*, 99755.

Fernandez, J., Wehrstedt, L., Shamis, L., Elhoushi, M., Saladi, K., Bisk, Y., et al. (2024). Hardware scaling trends and diminishing returns in large-scale distributed training. arXiv preprint arXiv:241113055.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, E., et al. (2023). "I wouldn't say offensive but…": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 205–216).

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024) *Bias and fairness in large language models: A survey.* https://arxiv.org/abs/2309.00770

Glazko, K. S., Yamagami, M., Desai, A., Mack, K. A., Potluri, V., Xu, X., et al. (2023). An autoethnographic case study of generative artificial intelligence's utility for accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 1–8).

Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., & Roberts, D. A. (2024). *The unreasonable ineffectiveness of the deeper layers*. https://arxiv.org/abs/240317887

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., et al. (2024). A survey on LLM-as-a-judge. arXiv preprint arXiv:241115594.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., et al. (2024) Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:240201680.

Hoadley, C., & Vogel, S. (2024). Autocorrect is not: People are multilingual and computer science should be too. *Communications of the ACM., 67*(2), 16–18.

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). Sage.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., et al. (2024). Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1395–1417).

Jacucci, G., Daee, P., Vuong, T., Andolina, S., Klouche, K., Sjöberg, M., et al. (2021). Entity recommendation for everyday digital tasks. *ACM Transactions on Computer-Human Interaction (TOCHI), 28*(5), 1–41.

Kaffee, L. A., Atanasova, P., & Rogers, A. (2025). Local differences, global lessons: Insights from organisation policies for international legislation. arXiv preprint arXiv:250305737.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. arXiv preprint arXiv:210314659.

Kharchenko, J., Roosta, T., Chadha, A., & Shah, C. (2024). How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. arXiv preprint arXiv:240614805.

Kibriya, H., Khan, W. Z., Siddiqa, A., & Khan, M. K. (2024). Privacy issues in large language models: A survey. *Computers and Electrical Engineering., 120*, 109698.

Kim, T. S., Lee, Y., Shin, J., Kim, Y. H., & Kim, J. (2024). Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–21).

Klouche, K., Ruotsalo, T., Cabral, D., Andolina, S., Bellucci, A., & Jacucci, G. (2015). Designing for exploratory search on touch devices. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 4189–4198).

Kumar, D., Jain, U., Agarwal, S., & Harshangi, P. (2024a). Investigating implicit bias in large language models: A large-scale study of over 50 LLMs. arXiv preprint arXiv:241012864.

Kumar, R., Vinisha, V., & Rajkumar, M. (2024b). A study on implementation of artificial intelligence technologies in business automachine. *International Journal of Advanced Research in Commerce, Management & Social Science, 7*, 206–210.

Kwan, H. Y. (2024). User-focused telehealth powered by LLMs: Bridging the gap between technology and human-centric care delivery. In *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)* (pp. 187–191). IEEE.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K. R., et al. (2024). RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, et al. (Eds.), *Proceedings of the 41st International Conference on Machine Learning, vol. 235 of Proceedings of Machine Learning Research PMLR* (pp. 26874–26901). https://proceedings.mlr.press/v235/lee24t.html

Lee, H. P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., et al. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems ACM*.

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024a). Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems, 37*, 84799–84838.

Li, Q., Hong, J., Xie, C., Tan, J., Xin, R., Hou, J., et al. (2024b). Llm-pbe: Assessing data privacy in large language models. arXiv preprint arXiv:240812787.

Li, Y., Zhang, L., Zhang, Y. (2024c). *Fairness of ChatGPT*. https://arxiv.org/abs/2305.18569

Liang, X., Wang, H., Wang, Y., Song, S., Yang, J., Niu, S., et al. (2024). Controllable text generation for large language models: A survey. arXiv preprint arXiv:240812599.

Liao, Q. V., & Wortman Vaughan, J. (2024). AI transparency in the age of LLMs: A human centered research roadmap. *Harvard Data Science Review*. (Special Issue 5). https://hdsr.mitpress.mit.edu/pub/aelql9qy/release/2

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–15).

Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., et al. (2023a). Aligning large language models with human preferences through representation engineering. arXiv preprint arXiv:231215997.

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Guo, R., Cheng, H., et al. (2023b). Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. arXiv preprint arXiv:230805374.

Masson, D., Malacria, S., Casiez, G., & Vogel, D. (2024). DirectGPT: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems CHI '24*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642462

McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 15458–15463).

Mellatdoust, P., Di Fede, G., Alrabie, L., Cappiello, C., & Andolina S. (2025). BiasAlert: Supporting Bias Identification in Search Queries. In *16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), Salerno, Italy*. ACM, New York, NY, USA. https://doi.org/10.1145/3750069.3755952

Meskó, B. (2023). The impact of multimodal large language models on health care's future. *Journal of Medical Internet Research, 25*, e52865.

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3470–3487). Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.244/

Mondal, H., & Mondal, S. (2024). Ethical and social issues related to AI in healthcare. *Methods in Microbiology, 55*, 247–281.

Morales, S., Clarisó, R., & Cabot, J. (2024). A DSL for testing LLMs for fairness and bias. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems* (pp. 203–213).

Morrison, J., Na, C., Fernandez, J., Dettmers, T., Strubell, E., & Dodge, J.. (2025). Holistically evaluating the environmental impact of creating language models. arXiv preprint arXiv:250305804.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality, 15*(2), 1–21.

Norman, D. A. (2013). *The design of everyday things*. MIT Press.

Nov, O., Singh, N., & Mann, D. (2023). Putting ChatGPT's medical advice to the (Turing) test: Survey study. *JMIR Medical Education, 9*, e46939.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., et al. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction, 39*(3), 391–437.

Pardo Gutierrez, S., Alrabie, L., Di Fede, G., Vitali, M., & Andolina, S. (2025). A Direct manipulation interface for LLM-based process modeling. In *16th biannual conference of the italian SIGCHI Chapter (CHItaly 2025), Salerno, Italy*. ACM, New York, NY, USA. https://doi.org/10.1145/3750069.3755960

Pathak, M., Shah, B., & Thakkar, A. (2025). Unveiling the potential: Large language models in financial sentiment analysis, education, and market analysis. In *AIP conference proceedings* (Vol. 3255). AIP Publishing.

Pucci, E., Piro, L., Andolina, S., & Matera, M. (2024). From conversational web to inclusive conversations with LLMs. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces AVI '24*. Association for Computing Machinery. https://doi.org/10.1145/3656650.3656739

Pyae, A. (2025). The human-AI handshake framework: A bidirectional approach to human-AI collaboration. arXiv preprint arXiv:250201493.

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), 27.

Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR), 55*(2), 1–39.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., et al. (2023). From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1–9). IEEE.

Samuelson, P. (2025). California's AI act vetoed. *Communications of the ACM, 68*(3), 18–20. https://doi.org/10.1145/3710808

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:170706347.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., et al. (2024). Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions. arXiv preprint arXiv:240609264.

Shneiderman, B. (1984). The future of interactive systems and the emergence of direct manipulation. In *Proceedings of the NYU Symposium on User Interfaces on Human Factors and Interactive Computer Systems USA* (pp. 1–28). Ablex Publishing Corp.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504.

Singh, A., Patel, N. P., Ehtesham, A., Kumar, S., & Khoei, T. T. (2024). A survey of sustainability in large language models: Applications, economics, and challenges. arXiv preprint arXiv:241204782.

Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., & Seifert, C. (2024). Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–19).

Sun, Q., Li, Y., Alturki, E., Murthy, S. M. K., Schuller, B. W. (2024). Towards friendly AI: A comprehensive review and new perspectives on human-AI alignment. arXiv preprint arXiv:241215114.

Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Hannay, S. B. W., et al. (2025). Ambient artificial intelligence scribes: Learnings after 1 year and over 2.5 million uses. *NEJM Catalyst, 6*(5), CAT.25.0040. https://doi.org/10.1056/CAT.25.0040

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288.

Valencia, S., Cave, R., Kallarackal, K., Seaver, K., Terry, M., & Kane, S. K. (2023). "The less I type, the better": How AI language models can enhance or impede communication for AAC users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Veturi, S., Vaichal, S., Jagadheesh, R. L., Tripto, N. I., & Yan, N. (2024). Rag based question-answering for contextual response prediction system. arXiv preprint arXiv:240903708.

Vuong, T., Andolina, S., Jacucci, G., & Ruotsalo, T. (2021a). Does more context help? Effects of context window and application source on retrieval performance. *ACM Transactions on Information Systems (TOIS), 40*(2), 1–40.

Vuong, T., Andolina, S., Jacucci, G., & Ruotsalo, T. (2021b). Spoken conversational context improves query auto-completion in web search. *ACM Transactions on Information Systems (TOIS), 39*(3), 1–32.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., et al. (2023a). Self-instruct: Aligning language models with self-generated instructions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13484–13508). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.754/

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., et al. (2023b) *Aligning large language models with human: A survey*. https://arxiv.org/abs/2307.12966

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345.

Warr, M., Oster, N. J., & Isaac, R. (2024). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education.*, 1–24.

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design principles for generative AI applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–22).

Whitmore, S., Harrington, C., & Pritchard, E. (2024). Assessing the ineffectiveness of synthetic reinforcement learning feedback in fine-tuning large language models. *OSF Preprints*. osf.io/cvdzu v1

Wu, T., Terry, M., & Cai, C. J. (2022). 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* (pp. 1–22).

Xiao, Z., Deng, W. H., Lam, M. S., Eslami, M., Kim, J., Lee, M., et al. (2024). Human-centered evaluation and auditing of language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–6).

Xiaoliang, M., RuQiang, Z., Ying, L., Congjian, D., & Dequan, D. (2024). Design of a large language model for improving customer service in telecom operators. *Electronics Letters, 60*(10), e13218.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., et al. (2023). Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems, 36*, 69798–69818.

Xu, J. (2024). GenAI and LLM for financial institutions: A corporate strategic survey. Available at SSRN 4988118.

Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6268–6278). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.385/

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., et al. (2024a). On protecting the data privacy of large language models (LLMs): A survey. arXiv preprint arXiv:240305156.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2024b). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112.

Yang, C., Xu, C., & Qi, Y. (2024). Financial knowledge large language model. arXiv preprint arXiv:240700365.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems CHI '23*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388

Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., et al. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys, 57*(5). https://doi.org/10.1145/3711118

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). 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 FAT* '20* (pp. 295–305). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. arXiv preprint arXiv:230318223, *1*(2).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024a). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–38.

Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., et al. (2024b). Revolutionizing finance with llms: An overview of applications and insights. arXiv preprint arXiv:240111641.

Zheng, Y., Koh, H. Y., Yang, M., Li, L., May, L. T., Webb, G. I., et al. (2024). Large language models in drug discovery and development: From disease mechanisms to clinical trials. arXiv preprint arXiv:240904481.

Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al.: Fine-tuning language models from human preferences; 2020. https://arxiv.org/abs/1909.08593.