A Survey of LLMs Responsible AI and Cultural Considerations

www.surveyx.cn

Abstract

This survey paper explores the intersection of Large Language Models (LLMs), Responsible AI, and cultural considerations, emphasizing their critical role in addressing cultural bias and localization challenges. In domains like healthcare and education, LLMs have shown promise in enhancing service delivery, yet they require robust evaluation methods and human oversight to mitigate ethical concerns such as data privacy and bias. Despite advancements, traditional LLM approaches face reliability issues in tasks requiring personalization, highlighting the need for comprehensive datasets that include diverse persona attributes. Progress in automated evaluation and self-correction capabilities has improved AI system reliability, but challenges like benchmark leakage and the need for dynamic dataset updates persist. Future research should expand datasets, refine evaluation methods, and focus on high-quality datasets for low-resource languages. Efforts should also enhance training methods for Multimodal Large Language Models (MMLLMs) and explore domain-specific adaptations to improve medical AI applications. By advancing these areas, responsible AI can evolve to ensure systems are culturally sensitive, ethically sound, and aligned with global needs. Key takeaways suggest integrating LLMs with causal discovery methods, monitoring responsible AI evolution, and leveraging advanced AI techniques for enhanced understanding. Addressing biases related to gender, race, and physical ability is crucial for developing inclusive datasets. A robust framework for trustworthy AI is essential for fostering public trust and ensuring AI benefits society while minimizing risks. Future research should refine ethical frameworks, explore LLM biases, and develop reliable ethical reasoning methods.

1 Introduction

1.1 Significance of Responsible AI

Responsible AI is crucial in the realm of large language models (LLMs), given the ethical and cultural challenges they present, including privacy, fairness, hallucination, and accountability. Addressing these issues requires tailored ethical frameworks and interdisciplinary collaboration to mitigate biases and enhance transparency in information dissemination [1, 2]. As LLMs integrate into critical sectors like healthcare, education, and employment, the development of comprehensive frameworks that uphold ethical standards and minimize cultural biases becomes essential for aligning AI systems with societal values and promoting inclusivity.

The foundational principles of responsible AI—fairness, robustness, transparency, accountability, privacy, and safety—are vital for preventing discriminatory outcomes and fostering trust in AI technologies [3]. These principles address privacy and fairness while ensuring the integrity of AI outputs by minimizing phenomena such as 'hallucination'. In the educational sector, it is imperative to equip practitioners with the knowledge to implement responsible AI principles effectively [4], bridging the gap between ethical principles and actionable practices [5].

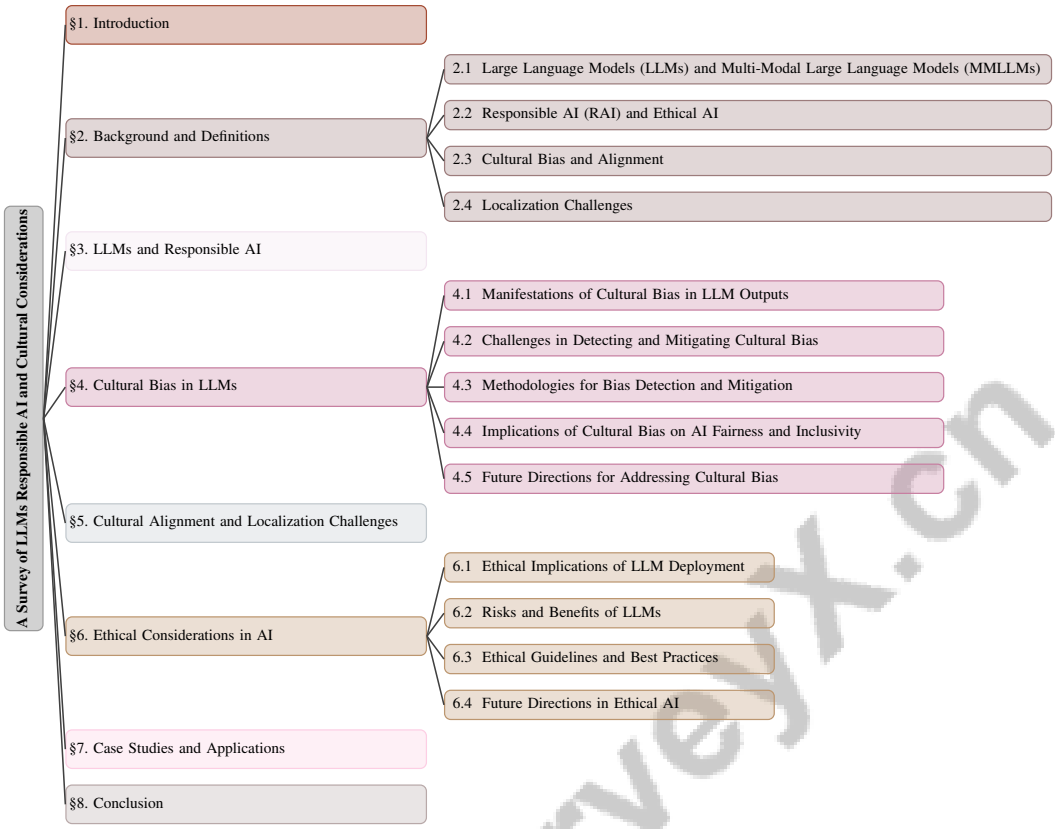

Figure 1: chapter structure

Cultural considerations are also pivotal, as responsible AI must navigate the challenges of interacting with diverse linguistic and cultural groups. Understanding the impact of sociodemographic factors on model behavior is essential for ensuring responsible AI use and cultural sensitivity [6]. The development and deployment of AI technologies, particularly LLMs, necessitate concerted efforts to address cultural biases and localization challenges, ensuring inclusivity and adherence to ethical standards that respect universal human rights [1, 7]. The significance of responsible AI lies in its potential to enhance fairness, accountability, and transparency, fostering trust among users from diverse backgrounds.

1.2 Role of Large Language Models

Large Language Models (LLMs) play a critical role in navigating the cultural and ethical challenges associated with AI systems. Their extensive linguistic capabilities allow them to operate across various cultural contexts, addressing issues of cultural sensitivity and ethical compliance [8]. By integrating ethical principles with regulatory requirements, LLMs support responsible AI development that aligns with societal norms [9].

In sectors such as healthcare and finance, the deployment of LLMs relies on effective methodologies to detect and mitigate harmful outputs, which is vital for maintaining ethical compliance and user trust [8]. Their ability to simulate diverse stakeholder perspectives enhances inclusivity and facilitates participatory decision-making processes [10]. This capability is particularly beneficial in education, where LLMs can generate tailored content to meet the unique needs of learners, including those who are Deaf and Hard of Hearing (DHH) [11].

LLMs also contribute to cultural analytics by aiding the classification and interpretation of cultural norms [12]. The evolution of LLMs and Multimodal Large Language Models (MLLMs) in medical practice highlights their role in integrating diverse data types, which is crucial for addressing cultural and ethical issues [13]. Nonetheless, challenges remain in processing low-resource and non-English

AI-generated, for reference only.

languages, underscoring the need for models that enhance multilingual capabilities and bridge gaps in natural language processing [14].

The sensitivity of LLMs to prompt designs can lead to skewed outcomes, emphasizing their role in addressing cultural and ethical challenges [15]. By fostering collaboration between humans and LLMs, these models can improve factual accuracy and source attribution, enhancing transparency in AI systems [16]. As AI evolves, aligning LLMs with ethical guidelines remains critical to ensuring culturally sensitive and ethically sound AI systems [17]. Additionally, preference alignment is vital for enhancing MLLM performance, particularly in reducing hallucinations and improving accuracy in image understanding tasks [18]. Unbiased watermarking methods further support the integration of ethical principles without compromising text quality [19].

1.3 Objectives of the Survey

This survey aims to provide a comprehensive overview of the current research landscape regarding the integration of Large Language Models (LLMs) across various domains, particularly focusing on responsible and ethical AI practices [20]. A primary objective is to establish actionable metrics for AI risk management that operationalize accountability and ensure adherence to ethical standards [21]. This involves examining methodologies for embedding responsible AI principles into system-level design patterns, facilitating the co-production of Responsible AI (RAI) values in AI/ML development.

The survey also explores the feasibility of incorporating unbiased watermarking techniques in LLMs, demonstrating their implementation without compromising output quality [19]. This addresses challenges in maintaining output integrity while embedding ethical considerations in AI systems. Additionally, the survey investigates the socio-technical contexts of AI deployment, emphasizing the need for frameworks that support both technical and non-technical stakeholders in AI application development.

By synthesizing insights from existing literature, the survey aims to provide policy recommendations that enhance LLM multilingual capabilities, addressing risks associated with multilingual jailbreaks and improving proficiency in low-resource languages. This comprehensive approach seeks to foster responsible AI practices that are culturally sensitive and ethically sound, reflecting diverse user perspectives. Engaging the public in AI governance and development promotes equitable practices and fosters user trust. The findings will offer valuable insights for developers and policymakers, helping them recognize individual and group-level cultural perspectives, ultimately addressing societal concerns and ensuring AI technologies are designed with fairness, accountability, and transparency [22, 1, 23, 24].

1.4 Structure of the Survey

This survey is systematically organized to comprehensively examine the interplay between Large Language Models (LLMs), responsible AI, and cultural considerations. The paper's structure is delineated into key sections, each addressing distinct aspects of the topic while maintaining thematic coherence.

The survey begins with an introduction that underscores the significance of responsible AI in the context of LLMs, followed by an exploration of the role these models play in addressing cultural and ethical challenges. The introduction also outlines the main objectives of the survey, setting the stage for subsequent sections.

The background section provides foundational definitions and explanations of key concepts such as LLMs, responsible AI, cultural bias, cultural alignment, localization challenges, and ethical AI. This section establishes a common understanding of the terms and their relevance to AI development and deployment.

The survey then delves into the role of LLMs in promoting responsible AI practices, highlighting existing frameworks and guidelines. This is followed by a detailed examination of cultural bias in LLMs, where methodologies for bias detection and mitigation are organized into four stages: pre-processing, in-training, intra-processing, and post-processing [25].

Subsequent sections analyze the challenges of cultural alignment and localization, emphasizing the importance of adapting AI systems to diverse cultural contexts. Strategies for overcoming localization

challenges are discussed, drawing on emerging frameworks for contextual grounding in textual models to improve ethical alignment [26].

The ethical considerations section explores the implications of deploying LLMs across various applications, discussing potential risks and benefits, and highlighting ethical guidelines and best practices. This section is organized around six primary themes: governance, fairness, explainability, human flourishing, privacy, and security, reflecting their interdependent nature in AI contexts [24].

The survey includes case studies and applications of LLMs that exemplify strategies for addressing cultural bias and localization challenges, highlighting their effectiveness and limitations in translating low-resource languages, evaluating and mitigating biases, and navigating ethical dilemmas unique to LLMs. This comprehensive overview aims to foster equitable AI technologies [27, 28, 2]. The survey concludes with a synthesis of key findings and suggestions for future research directions, contributing to the development of responsible AI practices that are culturally sensitive and ethically sound.

The survey organizes current methods into three stages: prevalent practices (reactive), emerging practices (proactive), and aspirational future practices (anticipatory), providing a roadmap for advancing responsible AI in the context of LLMs [29].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs) and Multi-Modal Large Language Models (MMLLMs)

Large Language Models (LLMs), based on the Transformer architecture, represent a significant leap in AI, adept at multilingual text processing and generation, thus playing a pivotal role in strategic domains [6]. Their application in complex environments, exemplified by Richelieu's capabilities in diplomacy and negotiation, demonstrates their utility in enhancing decision-making processes [30]. LLMs are also pivotal in sectors like healthcare and education, offering clinical decision support and personalized learning materials, including tailored content for Deaf and Hard of Hearing learners, thereby enhancing educational experiences [11, 31]. Techniques like CALM further augment LLMs' performance, highlighting their adaptability across sectors [32, 33].

The integration of LLMs into AI systems aligns with local ethical standards, fostering trust among diverse users. However, their computational demands necessitate optimization for efficient performance [6]. Multi-Modal Large Language Models (MMLLMs) extend LLM capabilities by integrating text, images, and audio, crucial for tasks like clinical decision-making and patient engagement in healthcare [13]. The development of alignment algorithms enhances MMLLMs' processing capabilities, supporting applications like text-to-video generation and image captioning across diverse domains [34, 35]. Continuous research aims to refine these models, ensuring they remain ethically aligned and resource-efficient [33].

2.2 Responsible AI (RAI) and Ethical AI

Responsible AI (RAI) and Ethical AI frameworks are integral to aligning AI systems with societal values and ethical norms, emphasizing principles like fairness, privacy, and accountability to foster user trust [36]. Implementing these principles requires bridging the gap between high-level ethics and practical application, especially in software engineering [37]. Frameworks that guide AI's sustainability impacts ensure alignment with broader societal goals [38], crucial in sectors like healthcare where accuracy and bias management are vital [13]. Developing metrics catalogs operationalizes accountability within AI systems, enhancing performance evaluation [21]. The Responsible AI by Design methodology integrates ethical principles from the outset of AI development [3].

Research into ethical AI governance has proposed frameworks that improve the understanding and application of responsible AI practices [39]. These frameworks help practitioners identify AI risks and optimize LLMs for performance and interpretability, exemplifying responsible AI principles [6]. Addressing responsible AI challenges enhances the trustworthiness and acceptance of AI technologies, ensuring fair, accountable, and transparent deployment [5].

2.3 Cultural Bias and Alignment

Cultural bias in AI, particularly LLMs, stems from training datasets that reflect gender, race, and appearance biases, often rooted in Western-centric paradigms [40, 24]. Addressing this bias is complicated by varied trustworthiness interpretations across sectors [9]. LLMs' limited multilingual proficiency, especially in low-resource languages, highlights the need for balanced training data and adaptation strategies [41]. Benchmarks for evaluating culturally conditioned prompts are critical for identifying biases and measuring political bias in LLMs [42].

Cultural alignment ensures AI systems resonate with diverse users' norms and values, fostering trust. However, challenges in detecting hallucinations and aligning models with user-specified knowledge bases persist [43]. In education, biases in training data hinder AI's effectiveness for diverse learners, including Deaf and Hard of Hearing students, requiring inclusive and equitable AI design [11, 44]. Addressing these biases necessitates understanding cultural and linguistic attributes influencing AI interactions, ensuring AI systems are culturally and ethically aligned.

2.4 Localization Challenges

Localizing AI systems to diverse cultural contexts involves aligning with linguistic and cultural nuances, facing challenges like low-resource language knowledge gaps and performance disparities [45]. LLMs struggle with non-English text, affecting their ability to produce culturally relevant outputs [46]. Biases from training data and societal norms complicate localization, lacking metrics for assessing LLM alignment [47, 48].

In domains like telecommunications and cybersecurity, LLMs must adapt to local contexts for tasks like vulnerability detection, where cultural nuances are crucial [49]. Techniques like the Guide-Align approach improve model alignment with human values [50]. In journalism and healthcare, localization ensures contextually relevant outputs, requiring accurate cross-lingual knowledge transfer [51, 28].

Overcoming localization challenges involves developing frameworks that address responsible AI practices across contexts, enhancing cultural alignment and promoting equitable outcomes [22, 52, 1]. Effective localization ensures AI systems are culturally sensitive and effective across societal landscapes [24].

In recent years, the discourse surrounding Large Language Models (LLMs) has increasingly focused on the principles of Responsible AI. To elucidate this complex landscape, Figure 2 illustrates the hierarchical structure of key concepts related to LLMs and Responsible AI. This figure emphasizes the interconnections between frameworks and competencies, evaluation and benchmarking, and strategic frameworks. By categorizing the primary ideas alongside their respective subcategories, it highlights not only the methodologies employed in this field but also the challenges and future directions that are critical for the responsible development of AI technologies. Such a visual representation serves to enhance our understanding of the multifaceted nature of Responsible AI, providing a comprehensive overview that complements our ongoing analysis.

3 LLMs and Responsible AI

3.1 Frameworks and Competency for Responsible AI

Establishing frameworks and competencies for responsible AI is vital to ensure AI systems adhere to ethical standards and societal values. These frameworks encompass ethical, legal, and technical dimensions, promoting fairness, transparency, and accountability [4]. Methodologies like manual coding of literature facilitate the creation of actionable guidelines through expert contributions, reinforcing responsible AI practices. Alignment strategies are crucial for mitigating hallucinations and enhancing the reliability of Multimodal Large Language Models (MLLMs) [18]. Unbiased watermarking methods, such as -reweight and -reweight, preserve AI output quality while implementing effective watermarking solutions [19].

The ExploreGen framework demonstrates the use of LLMs to generate diverse AI use cases, classify risks, and promote responsible AI development [53]. This reflects a broader commitment to enhancing AI's ethical dimensions in line with societal norms. Current research categorizes responsible AI into

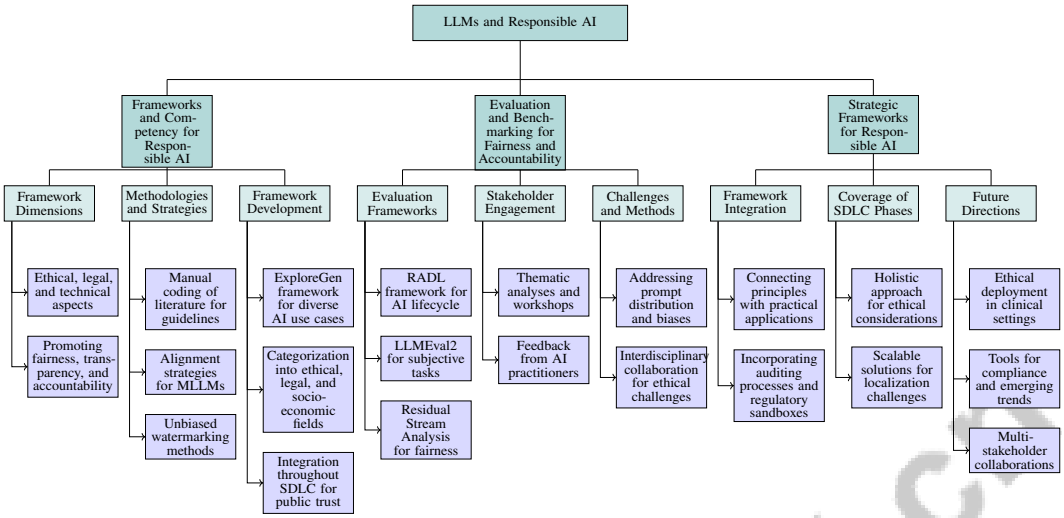

Figure 2: This figure illustrates the hierarchical structure of key concepts related to LLMs and Responsible AI, emphasizing frameworks and competencies, evaluation and benchmarking, and strategic frameworks. It categorizes the primary ideas and subcategories, highlighting methodologies, challenges, and future directions in responsible AI development.

fields like ethical AI, legal AI, and socio-economic considerations, emphasizing diverse stakeholder input and interdisciplinary dialogue [5]. This categorization is essential for developing comprehensive frameworks that address responsible AI's multifaceted challenges, offering practical guidance to enhance AI's trustworthiness and societal acceptance.

As illustrated in Figure 3, the key components of frameworks and competencies for responsible AI are categorized into ethical dimensions, technical strategies, and legal considerations. This figure highlights the integration of societal values, technical methodologies, and regulatory compliance, which are crucial for the development and deployment of responsible AI systems. Leveraging comprehensive frameworks improves AI's trustworthiness and societal acceptance, ensuring contributions to society align with ethical norms. Research highlights governance, fairness, and explainability while calling for exploration of privacy, security, and human flourishing. Integrating responsible AI principles throughout the Software Development Life Cycle (SDLC) fosters public trust and facilitates effective AI implementation in real-world contexts [54, 55, 24].

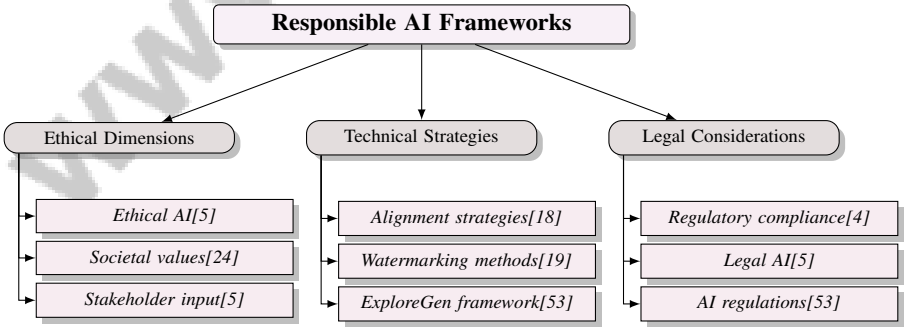

Figure 3: This figure illustrates the key components of frameworks and competencies for responsible AI, categorized into ethical dimensions, technical strategies, and legal considerations, highlighting the integration of societal values, technical methodologies, and regulatory compliance.

3.2 Evaluation and Benchmarking for Fairness and Accountability

Evaluating and benchmarking AI systems for fairness and accountability is crucial for responsible AI development, requiring robust frameworks that align with diverse societal values. Current studies often lack practical compliance mechanisms and fail to encompass the full spectrum of societal values

Benchmark	Size	Domain	Task Format	Metric
LLMEval2[56]	2,553	Text Summarization	Question Answering	Kappa correlation coefficient, Accuracy
LLM-Bench[57]	100,000	Natural Language Inference	Question Answering	Accuracy, ROUGE
TrustScore[58]	1,000	Question Answering	Open-ended Question Answering	TrustBC, TrustOV
ACR[59]	1,600	Cultural Bias Evaluation	Bias Detection And Jailbreaking	Attack Success Rate, Bias Detection Rate
RME[60]	2,000	Knowledge Editing	Factual Knowledge Editing	Accuracy, Reversion
LLM-Factuality[61]	2,250	Text Summarization	Factuality Evaluation	Partial Correlation Coefficient, Factuality Errors
LMDM[62]	5,000	Reading Comprehension	Multiple Choice Question Answering	Accuracy, Persuaded CAL-LLM[63]
100,000	Factual Generation	Multi-task Language Understanding	ECE, Reliability Diagram	

Table 1: The table presents a comprehensive overview of representative benchmarks used in the evaluation of large language models (LLMs) for fairness and accountability. It details the benchmark names, dataset sizes, domains, task formats, and evaluation metrics, providing a foundational understanding of the diverse methodologies employed in AI assessment. This information is crucial for understanding the multifaceted approaches to evaluating AI systems in alignment with societal values and ethical standards.

in AI design [64]. Addressing this gap requires comprehensive evaluation methods that incorporate diverse stakeholder perspectives and ethical considerations.

The RADL framework provides a structured approach to responsible AI development, incorporating stages such as Planning and Review, Design Review, Harm Modeling, Penetration Testing, and Incident Response to proactively mitigate potential harms [65]. This framework emphasizes integrating ethical considerations throughout the AI lifecycle, enhancing accountability and transparency.

Benchmarking tools like LLMEval2 assess LLM-generated responses in subjective and open-ended tasks, capturing human judgment nuances to promote fairness and accountability in AI outputs [56]. The Residual Stream Analysis method employs metrics such as accuracy, AUROC, and AUPRC to ensure AI systems meet fairness standards [66]. Performance assessments use accuracy metrics on held-out datasets, comparing composed models against baselines to evaluate improvements in task-specific capabilities [33]. Table 1 provides a detailed overview of various benchmarks utilized in the evaluation of AI systems, highlighting their relevance in ensuring fairness and accountability in AI development.

Thematic analyses of survey responses and workshops with AI practitioners, such as those at Meta, provide essential feedback for understanding responsible design patterns' practical implications [67]. These evaluations refine guidelines and ensure their relevance and usability, as evidenced by user studies involving AI researchers, engineers, designers, and product managers [4].

Specialized evaluation methods address challenges in evaluating LLMs amidst varying prompt distributions and inherent biases in benchmarks [57]. These methods enhance AI evaluations' robustness, ensuring benchmarks accurately reflect AI systems' performance and fairness. Frameworks categorizing roles such as researcher, data scientist, engineer, and policy analyst underscore the need for interdisciplinary collaboration to tackle ethical and technical challenges in AI systems [68].

Comprehensive evaluation and benchmarking methods are crucial for ensuring AI systems are fair, accountable, and aligned with societal values. Recent studies highlight AI technologies' ethical implications, particularly in text summarization, where stakeholder engagement and potential adverse impacts are often overlooked. Empirical research indicates that while AI ethics principles exist, they frequently lack practical guidance for developers, emphasizing the need to integrate ethical considerations throughout the AI lifecycle—from requirements engineering to deployment—to foster responsible AI development [69, 1]. Leveraging these methodologies enhances AI technologies' trustworthiness and societal acceptance, ensuring positive societal contributions while upholding ethical standards.

3.3 Strategic Frameworks for Responsible AI

Developing strategic frameworks for responsible AI is essential for guiding the ethical deployment of AI technologies, including Large Language Models (LLMs) and Multimodal Large Language Models

(MLLMs). These frameworks integrate theoretical principles with practical applications, ensuring AI systems align with societal values and ethical norms [9]. A significant innovation is the introduction of a comprehensive framework connecting these principles with practical applications, incorporating auditing processes and regulatory sandboxes to enhance AI systems' transparency and accountability.

A primary challenge in developing strategic frameworks is the insufficient coverage of Software Development Life Cycle (SDLC) phases by existing models, often focusing predominantly on the Requirements Elicitation phase [54]. Addressing this gap requires a holistic approach encompassing all SDLC phases, ensuring ethical considerations are integrated throughout the AI development process. This comprehensive coverage is vital for fostering a robust ethical foundation guiding AI systems' lifecycle.

Innovatively combining LLM capabilities with community-driven corpus development offers a scalable solution to localization challenges, especially in crises [70]. This approach exemplifies strategic frameworks' potential to enhance AI systems' cultural alignment and localization, ensuring responsiveness to diverse communities' specific needs. Additionally, a three-dimensional model—emphasizing transparency, integrity, and accountability—alongside AI Usage Cards represents a significant advancement in responsible AI practices, providing a structured methodology for evaluating and enhancing AI technologies' ethical deployment [71].

Future research should focus on developing robust methods for LLMs and MLLMs' ethical deployment in clinical settings, where stakes are particularly high [72]. Additionally, practical tools for compliance with ethical guidelines, exploration of emerging AI ethics trends, and fostering multi-stakeholder collaborations are necessary [64]. These efforts will advance strategic frameworks underpinning responsible AI, ensuring they address complex ethical challenges posed by rapidly evolving AI technologies.

Strategic frameworks for responsible AI are pivotal in guiding AI systems' ethical development and deployment. By integrating theoretical principles with practical applications and addressing the full SDLC spectrum, these frameworks ensure AI technologies are transparent, accountable, and aligned with societal values. Ongoing innovation and comprehensive research will enhance these frameworks' roles in promoting responsible and ethical AI application across sectors, addressing ethical considerations, mitigating biases, and ensuring AI systems' transparency and accountability. Despite the growing emphasis on responsible AI, many existing frameworks remain limited in scope, often focusing on early-stage development and lacking support for later SDLC phases. There is a pressing need for industry engagement in responsible AI research to bridge the gap between academic insights and practical applications, fostering trust and minimizing societal risks associated with AI technologies. Refining these frameworks to encompass a broader range of principles and provide actionable strategies will better support AI's diverse applications while safeguarding ethical standards [1, 73, 54, 23, 74].

4 Cultural Bias in LLMs

Understanding cultural bias in Large Language Models (LLMs) involves examining how biases in training datasets can distort outputs, leading to misrepresentations and disrespect for diverse cultural contexts. Table 2 offers a detailed comparison of different methodologies addressing cultural bias in Large Language Models, showcasing the manifestations of bias, challenges in detection and mitigation, and the strategies used for effective bias management. This section outlines specific examples of cultural bias in LLM outputs and their implications for accuracy and reliability in domains such as legal and educational settings, forming a basis for addressing broader challenges of cultural bias in AI systems.

4.1 Manifestations of Cultural Bias in LLM Outputs

Cultural bias in LLMs often originates from biases in their training datasets, resulting in outputs that inadequately respect diverse cultural contexts. For instance, LLMs exhibit varying accuracy in interpreting cultural norms, significantly influenced by the personas they adopt; models with socially favorable personas interpret norms more accurately [40]. In legal contexts, cultural bias challenges arise as LLMs must adapt to diverse regulatory environments, necessitating models that comprehend legal frameworks across regions like North America, India, and the UK [75]. Similarly, in educational

settings, cultural bias manifests when materials fail to accommodate the unique language capabilities of Deaf and Hard of Hearing (DHH) students, impacting the effectiveness of educational content [11].

Language generation by LLMs also reflects cultural bias, evident in biased language and toxic outputs, and in instances where models produce content unfaithful to input prompts [8]. Additionally, LLMs' personality traits can lead to biased interpretations influenced by cultural context [76]. The complexity and opacity of AI decision-making processes hinder accountability, obscuring the origins of bias [21]. This lack of transparency complicates efforts to trace and rectify biases, while optimization methods can indirectly affect fairness [6].

Addressing these manifestations is critical for developing strategies that enhance cultural sensitivity and inclusivity in AI systems, aligning technologies with ethical standards and fostering trust across diverse populations [22, 77, 7, 44, 24].

Figure 4 illustrates the manifestations of cultural bias in LLM outputs, categorizing them into cultural norms and personas, legal and educational bias, and language and personality influences. Each category highlights specific instances and research findings related to biases in large language models.

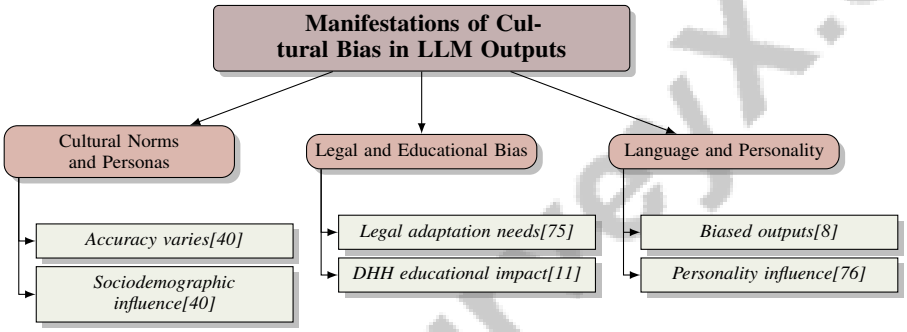

Figure 4: This figure illustrates the manifestations of cultural bias in LLM outputs, categorizing them into cultural norms and personas, legal and educational bias, and language and personality influences. Each category highlights specific instances and research findings related to biases in large language models.

4.2 Challenges in Detecting and Mitigating Cultural Bias

Detecting and mitigating cultural bias in LLMs is challenging due to the complexities of AI systems and their interactions with diverse cultural contexts. Core difficulties arise from inherent biases that distort cultural norm interpretations, complicating effective bias detection and mitigation [40]. Existing benchmarks often lack specificity and fail to measure nuanced harms in LLM outputs [8]. The 'black box' nature of AI models complicates efforts to promote responsible AI, as opacity and lack of consensus on ethical standards hinder transparency [78]. Furthermore, current responsible AI frameworks often overlook epistemological and political considerations shaping algorithmic impacts [79].

Benchmarks frequently neglect diverse cultural perspectives and ethical considerations crucial for accurate local value alignment, essential for detecting cultural bias [80]. In healthcare applications, biases in training datasets challenge bias detection and mitigation, highlighting the need for culturally sensitive approaches [13]. The scarcity of training data for non-English languages further complicates bias detection and mitigation, limiting models' adaptability to diverse linguistic contexts [46]. Additionally, Multimodal Large Language Models (MLLMs) often rely on inherent language biases rather than visual content, complicating evaluations of improvements in cultural bias mitigation [18].

LLMs' propensity to produce unrealistic or biased outputs necessitates additional validation and refinement processes to ensure cultural sensitivity and accuracy [53]. These challenges underscore the need for ongoing research and innovative methodologies to enhance the cultural inclusivity of AI systems, aligning them with ethical standards and fostering trust among diverse user groups.

4.3 Methodologies for Bias Detection and Mitigation

Detecting and mitigating cultural bias in LLMs is crucial for ensuring fairness and inclusivity across diverse cultural contexts. Various methodologies have been developed to address bias, encompassing evaluation methods categorized as data-level, model-level, and output-level, providing effective tools for bias identification. Mitigation strategies are classified into pre-model, intra-model, and post-model techniques, each with unique strengths and limitations. Recent studies emphasize structured guidelines and diverse data sources to enhance annotation consistency and reduce subjective variability [27, 1, 50].

Current methodologies often struggle to capture qualitative aspects, indicating a need for improved approaches to address cultural bias and alignment [81]. The LLMEval2 framework enhances assessment scope, being the largest and most diverse evaluation framework with 2,553 samples across 15 tasks and 8 abilities [56]. Systematic benchmarks introduce methods to test models against culturally sensitive and non-sensitive prompts, highlighting inconsistencies in model behavior across datasets [82].

Fine-tuning with Transfer Models (TMs) addresses performance issues in low-resource language contexts, providing insights into methodologies for detecting and mitigating cultural bias [83]. Integrating diverse data sources is crucial for effective failure management and bias detection in LLMs [84]. Methodologies like the hashing method, which substitutes specific words in prompts with unique identifiers, help prevent biases from affecting model responses [85].

The lack of standardization in definitions and slow policy adaptations to regulate AI technologies limit current research, underscoring the need for standardized benchmarks for bias assessment. Collaborative approaches among researchers are essential, as illustrated by the MULTITP dataset, which introduces systematic variations in moral dilemmas and supports over 100 languages [86]. Methodologies that detect low-resource knowledge queries and aggregate knowledge from other languages enhance LLM responses, improving cultural sensitivity and inclusivity [45]. The LLMs4OM framework exemplifies this by integrating retrieval with LLMs to improve ontology matching accuracy [87]. Additionally, emotional and visual aspects in question generation are critical for addressing learning difficulties faced by DHH students [11].

These methodologies reflect a commitment to employing innovative strategies and robust evaluation frameworks aimed at effectively identifying and addressing cultural biases in AI systems. This ensures that these systems not only demonstrate technical excellence but also adhere to ethical standards and promote cultural inclusivity. Recent literature emphasizes the importance of considering ethical implications and stakeholder impacts, particularly in sensitive applications like text summarization, healthcare, and employment [44, 1].

4.4 Implications of Cultural Bias on AI Fairness and Inclusivity

Cultural bias in LLMs poses significant challenges to AI fairness and inclusivity, as biases can lead to outputs that misrepresent or inadequately respect diverse cultural contexts. This misalignment undermines equitable AI deployment, impacting reliability and societal acceptance [38]. The absence of integrated frameworks encompassing all responsible AI principles across the Software Development Life Cycle (SDLC) complicates efforts to address cultural biases, limiting stakeholders' capacity to mitigate these biases effectively [4].

Cultural biases often compromise the statistical validity of LLM outputs, skewing predictive accuracy and affecting causal relations within training data [12]. The disconnect between trust and trustworthiness in AI systems highlights operational challenges in implementing AI ethics, directly affecting fairness and inclusivity [21]. Recognizing inherent biases in AI models emphasizes the necessity for ethical interventions, such as unbiased watermarking techniques, to track model outputs without compromising performance [19].

The fragmented nature of existing responsible AI practices often results in a lack of comprehensive applicability, perpetuating cultural biases [5]. The need for holistic governance approaches is evident, as many studies focus on procedural aspects of AI governance without adequately addressing stakeholder roles and timing [39]. This gap highlights the importance of developing frameworks that integrate technical and ethical considerations while engaging diverse stakeholders in AI development.

In educational contexts, while LLMs can enhance learning, they also pose risks regarding academic integrity and student understanding [20]. These risks are exacerbated by cultural biases influencing educational content generated by LLMs, affecting inclusivity in educational applications.

Addressing cultural bias in AI systems is crucial for fostering fairness and inclusivity, particularly as these technologies increasingly influence critical areas like healthcare, education, and employment. Given the potential for AI models to perpetuate discrimination against various demographics—including gender, ethnicity, and disability—developers must actively engage with diverse perspectives and implement robust bias mitigation strategies. Aligning AI development with universal human rights principles and incorporating public opinion insights ensures that AI technologies are equitable and responsive to global populations' needs. This comprehensive approach is vital for creating responsible AI that serves all societal members [44, 25, 7, 22]. By fostering trust and acceptance across cultural landscapes, AI systems can better fulfill their intended purposes, promoting equity and inclusivity in applications.

4.5 Future Directions for Addressing Cultural Bias

Addressing cultural bias in LLMs requires ongoing research and development to ensure models are fair, inclusive, and aligned with diverse cultural contexts. Future research should prioritize enhancing LLM interpretability, crucial for understanding decision-making processes and identifying potential biases in outputs [88]. Developing unified frameworks that address all pillars of Responsible AI (RAI) and improve interpretability of AI explanations is essential for fostering transparency and trust [78].

To effectively mitigate cultural bias, future research should focus on developing practical, context-sensitive fairness metrics applicable across various AI systems [44]. These metrics should enhance user experience and ensure responsiveness to the cultural and linguistic needs of diverse user groups. Additionally, exploring emerging technologies' implications for bias mitigation will be crucial in adapting to the evolving AI landscape.

Integrating diverse modalities in LLMs presents an opportunity to enhance cultural sensitivity and inclusivity. Future research should aim to develop robust methodologies for integrating these modalities, addressing challenges such as the hallucination problem through improved training techniques [88]. Furthermore, exploring sophisticated alignment techniques and enhancing dataset diversity will be critical for developing robust benchmarks to evaluate hallucinations in MLLMs [18].

Refining existing methodologies like the hashing technique and exploring automated identification of bias-inducing words can further contribute to bias mitigation efforts [85]. Additionally, expanding LLM applications to a broader range of pragmatic features and other speech acts beyond apologies will enhance their effectiveness across diverse cultural contexts [89].

Future research should prioritize comprehensive strategies that integrate innovative techniques and diverse datasets to address both intrinsic and extrinsic cultural biases in LLMs effectively. This includes employing structured annotation benchmarks for enhanced labeling consistency, utilizing multi-source data to improve model performance, and leveraging collaborative approaches among LLMs to identify and address knowledge gaps. By synthesizing these methodologies, researchers can work towards creating fairer and more responsible AI systems, fostering equitable technologies across applications such as healthcare and criminal justice [27, 90, 32, 50]. Addressing these challenges will enable AI systems to achieve greater cultural alignment, ensuring outputs are fair, inclusive, and effective across diverse cultural landscapes.

Feature	Manifestations of Cultural Bias in LLM Outputs	Challenges in Detecting and Mitigating Cultural Bias	Methodologies for Bias Detection and Mitigation
Bias Detection	Dataset Analysis	Complexity Issues	Data-level Methods
Mitigation Strategy	Cultural Sensitivity	Responsible AI Frameworks	Pre-model Techniques
Evaluation Framework	Not Specified	Lacks Specificity	Llmeval2 Framework

Table 2: This table provides a comprehensive comparison of methodologies for detecting and mitigating cultural bias in Large Language Models (LLMs). It highlights the manifestations of cultural bias in LLM outputs, the challenges in detecting and mitigating these biases, and the various methodologies employed for bias detection and mitigation. The table underscores the complexity of achieving cultural sensitivity and inclusivity in AI systems, emphasizing the need for robust evaluation frameworks and innovative strategies.

5 Cultural Alignment and Localization Challenges

5.1 Challenges of Cultural Alignment

Aligning Large Language Models (LLMs) with diverse cultural contexts is challenging due to the intricate nature of cultural nuances and current AI systems' limitations. A significant hurdle is the lack of understanding of low-resource languages, which hinders effective communication and cultural alignment [46]. This issue is exacerbated by knowledge transfer difficulties, leading to catastrophic forgetting and inadequate domain adaptation [91].

Integrating LLMs into industries such as telecommunications involves aligning AI systems with specific cultural and industry nuances. Frameworks like LLM-Modulo, which rely on external critics, add complexity to this process [40]. Furthermore, the disconnect between internal Responsible AI (RAI) policies and their application within Environmental, Social, and Governance (ESG) frameworks complicates cultural alignment [13].

Metacognitive myopia impairs LLMs' ability to align with diverse cultural contexts [80]. The absence of standardized qualitative assessment tools further complicates these challenges, as they are essential for evaluating the cultural sensitivity of AI outputs [79].

In education, cultural knowledge is crucial, as seen in frameworks categorizing AI tutors by their Deaf and Hard of Hearing (DHH) education backgrounds, highlighting the need for diverse cultural perspectives in AI [11]. The challenge of detecting offensive language in casual contexts underscores the need for improved methodologies to address cultural biases [92].

Cultural context is vital in moral decision-making, necessitating research to develop methodologies that enhance cultural sensitivity and inclusivity in AI [75]. Future research should integrate diverse perspectives into AI ethics, examining how cultural and demographic backgrounds influence ethical priorities [93].

Richelieu, an LLM-based agent, demonstrates the potential of self-evolving systems to adapt to diplomatic scenarios without human intervention, showing how continuous learning can improve cultural alignment [30].

As illustrated in Figure 5, the primary challenges in aligning LLMs with diverse cultural contexts can be categorized into language challenges, cultural integration, and cultural sensitivity. Each category highlights specific issues, such as low-resource language understanding, industry-specific cultural alignment, and the need for standardized assessment tools to evaluate cultural sensitivity. Addressing cultural alignment challenges requires innovative methodologies and frameworks to enhance LLMs' cultural sensitivity and inclusivity while tackling ethical complexities such as accountability and transparency [12, 2, 51]. Overcoming these challenges will enable AI systems to achieve greater cultural alignment, ensuring their outputs are relevant and effective across diverse cultural landscapes.

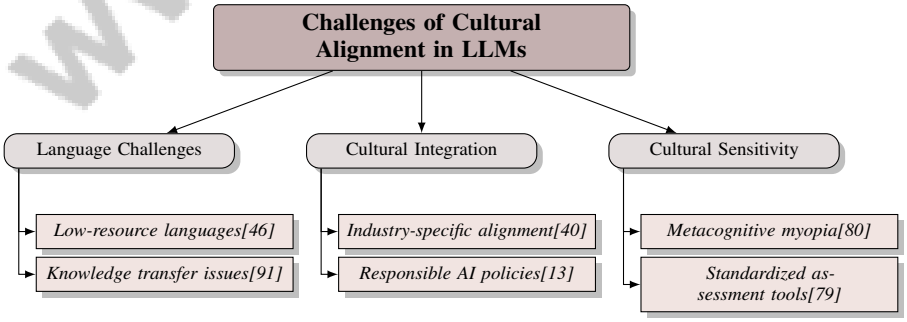

Figure 5: This figure illustrates the primary challenges in aligning Large Language Models (LLMs) with diverse cultural contexts, categorized into language challenges, cultural integration, and cultural sensitivity. Each category highlights specific issues, such as low-resource language understanding, industry-specific cultural alignment, and the need for standardized assessment tools to evaluate cultural sensitivity.

5.2 Localization in AI Systems

Localizing AI systems is crucial for adapting technologies to meet the linguistic, cultural, and contextual needs of diverse users, ensuring relevance and effectiveness. Localization provides accurate and contextually appropriate outputs through tailored translation processes and Transfer Models (TMs) [83], enhancing AI systems' applicability to specific organizational needs.

The complexity of localization is heightened by ethical considerations, requiring AI systems to align with local values and norms. Operationalizing ethical principles in localization is challenging due to AI's unique characteristics, necessitating nuanced approaches to address biases and ethical dilemmas [55]. This complexity is evident in educational applications, where LLMs must adapt to diverse grading standards and feedback mechanisms to ensure fairness and effectiveness [94].

A multidisciplinary approach is essential for addressing localization challenges. For instance, developing benchmarks for evaluating Chinese instruction understanding highlights the need to incorporate various categories for comprehensive localization [95]. Additionally, high computational and financial costs associated with LLMs pose significant challenges, necessitating innovative solutions to optimize resource usage [12].

In conversational AI, cost-effective personalization methods, such as reducing reliance on large datasets, illustrate the potential for localized AI systems to deliver tailored user experiences [96]. Future research should focus on establishing best practices for the ethical use of LLMs in education and other domains, examining their long-term impact on learning outcomes [20].

Localization of AI systems requires careful consideration of linguistic, cultural, and ethical factors. By addressing ethical and cultural challenges identified in responsible AI research, including studies on text summarization and human rights frameworks, AI systems can enhance cultural alignment and operational effectiveness. This approach ensures outputs are relevant and sensitive to global users' diverse needs, fostering equitable practices and mitigating adverse impacts across societal contexts [52, 1, 7, 22].

5.3 Strategies for Overcoming Localization Challenges

Overcoming localization challenges in AI systems requires a multifaceted approach that integrates diverse methodologies to enhance LLM adaptability across cultural contexts. Developing robust evaluation frameworks to assess LLM outputs is critical, focusing on bias detection and mitigation in knowledge engineering [97]. These frameworks ensure AI systems can sensitively respond to cultural and linguistic nuances.

Integrating multimodal data sources is vital for improving model performance and overcoming localization challenges [72]. By incorporating diverse modalities, AI systems can achieve a comprehensive understanding of cultural contexts, enhancing their ability to generate contextually relevant outputs. This approach is complemented by exploring self-correction capabilities and continual learning strategies, which can be expanded to multimodal settings to refine AI outputs [98].

Enhancing model versatility and improving dataset quality are key strategies for overcoming localization challenges. Retrieval-based approaches can complement generative processes, providing AI systems with the flexibility needed to adapt to diverse contexts [34]. Robust bias mitigation strategies, such as prompt debiasing filters and architectural innovations, can significantly reduce bias, improving AI outputs' cultural sensitivity [99].

Fine-tuning LLMs for specific annotation tasks and investigating different model architectures are strategies to enhance AI systems' localization [100]. These approaches enable AI systems to align better with local values and norms, ensuring outputs are relevant and effective across diverse cultural landscapes. Moreover, employing optimization algorithms can enhance AI adaptability in interdisciplinary applications, facilitating their integration into various contexts [6].

Addressing localization challenges involves leveraging innovative strategies and diverse datasets to enhance AI systems' cultural alignment and effectiveness. By proactively addressing cultural sensitivity and inclusivity challenges, AI systems can improve their relevance and effectiveness across diverse cultural contexts. This entails engaging stakeholders, understanding different user groups' unique needs, and continuously refining AI outputs to reflect evolving cultural dynamics. Such an

approach mitigates potential adverse impacts and promotes equitable practices in AI applications, ensuring technology serves a broad spectrum of societal interests and perspectives [52, 22, 1, 101, 51].

6 Ethical Considerations in AI

6.1 Ethical Implications of LLM Deployment

Deploying Large Language Models (LLMs) involves navigating significant ethical challenges to ensure alignment with societal norms. Key concerns include cognitive biases and overconfidence in AI outputs, potentially leading to misinterpretations in critical areas like healthcare and legal compliance. Establishing robust ethical frameworks is essential to guide LLM operations within accepted guidelines [8]. Recent advancements have enhanced understanding of LLMs' moral reasoning and biases [14]. Regulatory sandboxes provide secure environments for testing AI systems, facilitating ethical evaluations [9]. This approach is crucial for advancing AI safety and responsible deployment [8].

In healthcare, ethical concerns include data privacy due to sensitive patient information [13]. The risks of unfair bias and transparency deficits in algorithmic decision-making highlight the need for comprehensive Responsible AI (RAI) standards to mitigate potential harms. Ensuring equitable preference judgments in AI evaluations is critical for maintaining trust and transparency in LLM systems [42]. Moreover, LLMs' personalization capabilities present ethical risks, particularly regarding harmful persuasion, emphasizing the importance of transparency in AI personality shaping [76]. Current assessment techniques for LLM credences pose ethical challenges, necessitating improved methodologies for accurate AI output interpretation [102].

The U.S. Government's leadership in Responsible AI implementation can set standards and foster accountability and ethical integrity in AI deployment [17]. Findings from ExploreGen suggest enhancing responsible AI design by helping practitioners envision realistic AI applications and assess associated risks, contributing to ethical AI deployment [53]. Addressing ethical implications requires a comprehensive approach integrating transparency, accountability, and societal values, ensuring responsible LLM deployment and fostering trust among diverse user groups [19].

6.2 Risks and Benefits of LLMs

The deployment of Large Language Models (LLMs) presents a balance of risks and benefits that must be carefully managed for responsible and ethical use. Risks include potential privacy violations and data governance issues, as LLMs often rely on extensive datasets containing sensitive information [103]. The reliance on English-only interactions poses challenges for Deaf and Hard of Hearing (DHH) learners, exacerbating inequities in learning opportunities [11]. LLMs also exhibit limitations in assessing factual errors, with studies indicating a lack of correlation between model outputs and human evaluations for many error types [61]. This underscores the necessity for enhanced calibration techniques to improve trustworthiness, particularly in high-stakes applications [63]. Additionally, cybersecurity risks, including vulnerabilities to adversarial attacks, highlight the need for robust frameworks to guide ethical AI development [103].

Conversely, LLMs offer substantial benefits, particularly in their capacity to process and generate human-like text efficiently. The development of deeper and wider LLM networks has led to improvements in evaluation speed and cost-effectiveness, contributing to fairer assessments and more efficient AI applications [56]. Optimization algorithms enhance model performance and scalability, facilitating application across diverse domains [6].

To ensure responsible LLM deployment, a comprehensive strategy is essential, maximizing benefits such as improved text summarization and output quality while addressing risks like biased content generation and privacy concerns. Engaging relevant stakeholders and implementing safety-driven guidelines can enhance models' alignment with human values and ethical standards [49, 1]. By tackling challenges related to privacy, data governance, and factuality, and leveraging advancements in model calibration and optimization, LLMs can be developed and deployed in alignment with ethical standards and societal values, fostering trust among diverse user groups.

6.3 Ethical Guidelines and Best Practices

Establishing ethical guidelines and best practices is crucial for the responsible development and deployment of Large Language Models (LLMs), particularly in sensitive domains like healthcare. These guidelines ensure alignment with societal values and ethical standards, addressing the complexities of LLM integration [13]. A design-first approach is advocated, embedding responsible AI values throughout the development lifecycle, aligned with ISO standards and the EU AI Act [4].

Integrating ethical risk assessment tools into the development lifecycle is vital for identifying and mitigating potential concerns early in the process [37]. This proactive approach embeds ethical considerations as foundational elements of AI development, promoting transparency, accountability, and inclusivity.

In healthcare, specific ethical guidelines are necessary to address data privacy, patient safety, and fairness [13]. These guidelines provide frameworks for navigating ethical challenges, ensuring AI systems enhance patient care rather than compromise it.

Establishing comprehensive ethical guidelines and best practices is essential for building trust and ensuring societal acceptance of AI technologies, particularly regarding accountability, transparency, and privacy. Recent research emphasizes the need for actionable guidelines, informed by empirical studies that identify critical phases in AI design—from requirements engineering to deployment—capable of integrating ethical considerations effectively. By employing a multidisciplinary approach that incorporates techniques like differential privacy and federated learning, AI systems can respect individual privacy while aligning with ethical standards, fostering responsible AI use in society [104, 69]. By embedding ethical considerations throughout the AI development lifecycle and aligning with established standards, organizations can ensure fair, transparent, and ethically aligned AI deployments, enhancing system trustworthiness and positive contributions across various applications.

6.4 Future Directions in Ethical AI

Future directions in ethical AI should focus on developing adaptive practices and tools that bridge the gap between ethical guidelines and practical implementation. This includes creating methodologies that resolve tensions among ethical principles and exploring their interactions [105]. Expanding the scope of Human-Centered Ethical Research in AI (HCER-AI) to encompass privacy and security while examining societal impacts is crucial for advancing ethical AI practices [24]. Additionally, refining implementation frameworks and providing ongoing training and support for staff to adopt Responsible AI practices will be essential for embedding ethical considerations throughout the AI development lifecycle [17].

Research should also aim to integrate LLMs with specialized knowledge, developing long-form, open-ended questions to better assess contextual applications of knowledge [106]. Improving LLM calibration is necessary for enhancing the validity of statistical estimates, thus increasing the reliability of AI outputs [107]. Furthermore, exploring public opinion on responsible AI through qualitative methods can yield insights into societal attitudes, informing more inclusive and transparent AI policies [22].

As illustrated in Figure 6, future directions in ethical AI focus on three main areas: the development of methodologies to resolve ethical tensions and refine frameworks, the integration and improvement of large language models (LLMs), and the adoption of interdisciplinary approaches to develop regulatory frameworks and evaluation metrics. Interdisciplinary approaches addressing ethical AI challenges, promoting transparency, and developing adaptive regulatory frameworks should be a focus of future research [104]. This includes developing standardized evaluation metrics, improving alignment techniques, and exploring methodologies to enhance LLM trustworthiness [48]. Applying psychometric methods to LLM architectures could improve trait measurement accuracy, providing nuanced insights into AI behavior [76].

Additionally, research should focus on concrete methods for integrating Responsible AI (RAI) into corporate strategies, examining stakeholder roles, and exploring innovative approaches to overcome identified barriers [16]. Future efforts should also aim to refine the ESG Digital and Green Index framework for various contexts, suggesting directions for ethical AI practices [38]. Through these comprehensive research initiatives, the field of ethical AI can progress towards more transparent,

accountable, and culturally inclusive AI systems, fostering trust and acceptance across diverse user groups.

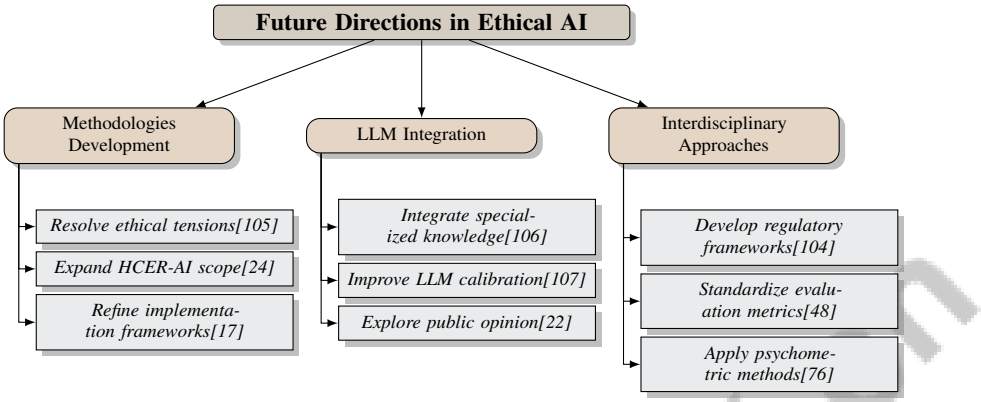

Figure 6: This figure illustrates the future directions in ethical AI, focusing on three main areas: the development of methodologies to resolve ethical tensions and refine frameworks, the integration and improvement of large language models (LLMs), and the adoption of interdisciplinary approaches to develop regulatory frameworks and evaluation metrics.

7 Case Studies and Applications

7.1 LLMs in Healthcare and Education

The integration of Large Language Models (LLMs) in healthcare and education offers significant opportunities to enhance service delivery and educational outcomes, while necessitating careful consideration of cultural and ethical dimensions. Figure 7 illustrates this integration, highlighting key applications and ethical considerations. In healthcare, LLMs assist in clinical decision-making by analyzing patient data to propose diagnoses or treatment plans, raising ethical concerns regarding data privacy, bias, and AI-generated recommendations' reliability [13]. This figure emphasizes the importance of addressing cultural sensitivity within healthcare applications, which is crucial for cultivating trust and acceptance among diverse demographic groups. To mitigate bias and enhance cultural sensitivity, it is essential that these models are trained on varied datasets.

In education, LLMs have the potential to transform learning experiences by personalizing educational content to meet the needs of diverse learners, including those from unique cultural backgrounds and students with specific learning challenges, such as Deaf and Hard of Hearing (DHH) individuals [11]. The figure also underscores how LLMs can significantly improve inclusivity and accessibility, thereby promoting equity in learning opportunities. However, deploying LLMs in education requires a cautious approach to address ethical concerns related to academic integrity and the risk of perpetuating biases through AI-generated content. Establishing ethical guidelines and best practices is essential for responsible utilization, fostering an educational environment that is inclusive and fair [20].

Both sectors require rigorous alignment with ethical standards and cultural norms to navigate the unique challenges posed by LLMs, including privacy, fairness, and accountability issues. Recent studies highlight challenges such as hallucination, verifiable accountability, and censorship, necessitating tailored ethical frameworks and dynamic auditing systems to combat biases and enhance transparency. Furthermore, the deployment of personalized LLMs, influenced by assigned personas, underscores the importance of maintaining consistent interpretations of culturally relevant social norms, as variations in persona can affect the model's understanding of these norms. Interdisciplinary collaboration is thus crucial to responsibly integrate LLMs across various contexts [40, 2]. By addressing these cultural and ethical aspects, LLMs can significantly enhance service delivery and educational outcomes, fostering trust and acceptance among diverse user groups.

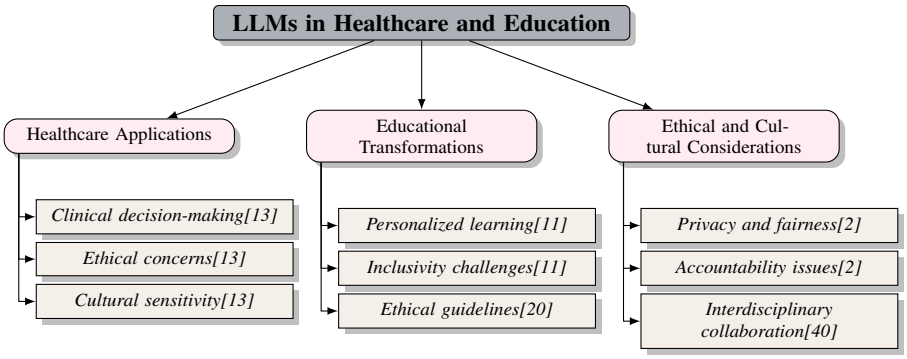

Figure 7: This figure illustrates the integration of Large Language Models (LLMs) in healthcare and education, highlighting key applications and ethical considerations. In healthcare, LLMs support clinical decision-making and address cultural sensitivity, while in education, they enhance personalized learning and inclusivity. The figure also emphasizes the importance of ethical and cultural considerations, including privacy, fairness, and accountability, underscoring the need for interdisciplinary collaboration.

7.2 Responsible AI in Cybersecurity and Media

The implementation of Responsible AI (RAI) in cybersecurity and media is critical for addressing the ethical and operational challenges inherent in these rapidly evolving fields. In cybersecurity, AI technologies enhance threat detection, vulnerability assessment, and incident response, providing advanced solutions to protect digital infrastructures. However, deploying AI necessitates adherence to responsible AI principles to mitigate risks associated with bias, privacy violations, and security breaches [33]. Transparency and accountability are essential for maintaining trust in cybersecurity applications, where the stakes are particularly high.

In media, AI plays a significant role in content generation, moderation, and distribution, raising ethical concerns about misinformation, bias, and public opinion manipulation. Implementing responsible AI frameworks is vital to uphold journalistic integrity and ensure accurate information dissemination [6]. Developing ethical guidelines and best practices is crucial for aligning AI-driven media applications with societal values, fostering trust and credibility among audiences.

Integrating responsible AI practices in both cybersecurity and media is imperative to effectively tackle the ethical and societal complexities inherent in these domains. This integration ensures that AI technologies are deployed ethically and transparently while prioritizing individual privacy and incorporating innovative algorithmic techniques such as differential privacy and federated learning. Ongoing research emphasizes governance, fairness, and explainability in AI systems, underscoring the need for a comprehensive approach that combines technological advancements with ethical and regulatory frameworks to mitigate risks and protect personal data [104, 24]. By addressing these challenges, AI systems can enhance the security and integrity of digital infrastructures and media applications, promoting trust and acceptance among diverse user groups.

7.3 Localization Challenges in Multilingual Models

The localization of multilingual AI models presents significant challenges due to the complexities of adapting these systems to diverse linguistic and cultural contexts. A core issue is the limited availability of high-quality training data in low-resource languages, which hampers the ability of Large Language Models (LLMs) to provide accurate and contextually relevant outputs across different languages [46]. This scarcity not only restricts the linguistic capabilities of AI systems but also exacerbates existing biases, as models trained predominantly on high-resource language data may overlook the nuances of less-represented languages [45].

Integrating domain-specific knowledge into multilingual models presents another challenge, as these systems must adapt to the unique requirements of various fields, such as telecommunications and healthcare [13]. This adaptation is complicated by the necessity to align AI outputs with local cultural norms and values, ensuring that models are both linguistically and culturally sensitive. Additionally,

the absence of standardized metrics for evaluating multilingual model performance across diverse languages and cultural contexts limits the ability to assess their effectiveness and trustworthiness [48].

Moreover, multilingual models must navigate local regulatory and ethical standards, which vary across regions and industries [9]. This alignment is crucial for ensuring responsible deployment of AI systems in compliance with local laws and guidelines. The complexity of these challenges underscores the need for innovative strategies and frameworks that enhance the localization capabilities of multilingual models, enabling effective and ethical operations across diverse cultural landscapes.

To address the localization challenges faced by multilingual AI models, a multifaceted strategy is essential, integrating diverse datasets, implementing robust evaluation metrics, and fostering interdisciplinary collaboration. Recent studies emphasize the importance of knowledge aggregation, fine-tuning on mixed-language data, and a comprehensive understanding of ethical considerations in AI development [41, 1, 108, 45]. By overcoming these challenges, AI systems can achieve greater cultural alignment and inclusivity, ensuring that outputs are relevant, sensitive, and effective across varied linguistic and cultural contexts.

7.4 Innovative Applications and Ethical AI

Innovative applications of Large Language Models (LLMs) that adhere to ethical AI principles are reshaping various domains by enhancing capabilities while ensuring alignment with societal values. The deployment of LLMs in multilingual contexts exemplifies their potential to bridge linguistic gaps and foster inclusivity. However, realizing this potential requires improved training strategies and the development of tailored architectures to enhance multilingual capabilities [41]. Integrating multilingual datasets is crucial to address linguistic diversity challenges and ensure that AI systems are culturally sensitive and contextually relevant.

In healthcare, LLMs are leveraged to provide personalized medical advice and support, improving patient outcomes by offering insights that are both accurate and culturally tailored. These applications highlight the necessity of ethical guidelines prioritizing patient privacy and data security, ensuring that AI technologies enhance healthcare delivery without compromising ethical standards. In education, LLMs transform learning experiences by generating adaptive materials that cater to diverse student populations, including marginalized groups like DHH learners. Research demonstrates that LLMs can create personalized quiz questions tailored to enhance video-based learning for DHH students, emphasizing visual and emotional learning strategies. Furthermore, LLMs are being fine-tuned to improve multilingual performance and inclusivity, ensuring accessibility for individuals from various linguistic backgrounds. These advancements underscore the potential of LLMs to foster equitable learning environments while emphasizing the importance of ethical considerations and fairness in their application [46, 11, 2, 109, 33]. This adaptability illustrates the role of LLMs in promoting equity and inclusivity in educational settings, aligning with ethical AI principles.

Moreover, the use of LLMs in creative industries, such as content creation and media, showcases their potential to generate innovative outputs while upholding ethical standards. The implementation of AI in applications like text summarization requires meticulous attention to critical issues such as bias and misinformation, highlighting the need for comprehensive frameworks that promote responsible AI usage. Recent research reveals a gap in addressing ethical implications and stakeholder engagement within the literature, as demonstrated by a study analyzing 333 summarization papers from 2020 to 2022, which found that few researchers adequately considered potential adverse impacts or engaged with relevant stakeholders. The introduction of AI Usage Cards aims to standardize reporting on AI applications, emphasizing principles of transparency, integrity, and accountability. Collectively, these efforts advocate for a more informed and ethical approach to deploying AI technologies in scientific research and beyond [1, 71]. By addressing these ethical considerations, LLMs can contribute to creating content that is innovative and aligned with societal values.

Exploring innovative applications of LLMs while adhering to ethical AI principles is crucial for unlocking their full potential across various domains, especially as we confront unique ethical challenges such as hallucination, accountability, and bias. Addressing these complexities through tailored ethical frameworks and interdisciplinary collaboration will enhance transparency and reduce biases, ensuring that LLMs positively contribute to information dissemination and scientific research practices. By focusing on practical, user-centered strategies that prioritize understanding model

outputs, respecting privacy, and promoting transparency, we can effectively integrate LLMs into diverse fields while upholding ethical standards and fostering responsible innovation [74, 2]. With improved training strategies, multilingual capabilities, and ethical guidelines, LLMs can be deployed in transformative and responsible ways, fostering trust and acceptance among diverse user groups.

8 Conclusion

This survey underscores the transformative potential of Large Language Models (LLMs) in addressing cultural biases and localization challenges, emphasizing their capacity to enhance service delivery while adhering to ethical standards. In domains like healthcare and education, LLMs significantly improve efficiency and knowledge dissemination, yet they necessitate rigorous evaluation and human oversight to mitigate ethical issues such as data privacy and bias. Despite advancements, traditional LLM frameworks, including models like LLM-as-a-Judge, still face challenges in personalization tasks, highlighting the need for comprehensive datasets that reflect diverse persona attributes.

Progress in developing automated evaluation and self-correction frameworks for LLMs has bolstered the reliability and validity of AI systems. However, issues like benchmark leakage and the need for dynamic dataset updates persist, calling for interdisciplinary research and robust governance frameworks to ensure AI technologies align with societal values and ethical norms. Future research should focus on expanding datasets and refining evaluation methods to enhance benchmark effectiveness for LLMs, alongside developing high-quality datasets for low-resource languages to improve training and governance structures.

Improving training methodologies for Multimodal Large Language Models (MMLLMs) and exploring domain-specific adaptations are crucial for enhancing the reliability and performance of medical AI applications. Advancing research in these areas is vital for the evolution of responsible AI, ensuring that AI systems remain culturally sensitive, ethically sound, and responsive to the diverse needs of global populations. Future studies should also delve into practical implementations of contestation processes and design systems that are user-friendly and effective.

Key insights suggest future research directions, including integrating LLMs with traditional causal discovery methods and examining the impact of context on causal inference. Monitoring the evolution of responsible AI, exploring interdisciplinary interactions, and leveraging advanced AI techniques for enhanced analysis and understanding are essential for future advancements. Employing LLMs as experimental agents can yield results comparable to human subjects in economic experiments, offering a promising avenue for exploration.

Future research should refine ethical frameworks for LLMs, examine the implications of their biases, and develop methods for more reliable ethical reasoning. Despite advancements in LLM-agent planning, challenges such as hallucinations and the feasibility of generated plans remain critical areas for further investigation. The CALM framework demonstrates substantial performance improvements across tasks by effectively integrating an anchor model with augmenting models, achieving capabilities unattainable by either model independently.

Addressing biases in LLMs related to gender, race, and physical ability is imperative, indicating future research directions for creating more inclusive datasets. Establishing a robust framework for trustworthy AI is essential to foster public trust and ensure AI technologies benefit society while minimizing associated risks. The experiments show that the new benchmark significantly enhances the ability to evaluate and improve the safety of LLMs, providing critical insights for future AI governance.

A crucial takeaway is that Explainable AI (XAI) is vital for developing trustworthy and socially responsible AI systems, facilitating understanding of model decisions and ensuring ethical compliance. Addressing these challenges will enable future research to contribute to healthier AI systems aligned with human values, fostering trust and acceptance among diverse user groups. The survey concludes that advancing responsible AI necessitates interdisciplinary collaboration and the establishment of best practices that conform to ethical guidelines.

References

[1] Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, and Adam Trischler. Responsible ai considerations in text summarization research: A review of current practices, 2023.

[2] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[3] Richard Benjamins, Alberto Barbado, and Daniel Sierra. Responsible ai by design in practice, 2019.

[4] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei. Rai guidelines: Method for generating responsible ai guidelines grounded in regulations and usable by (non-)technical roles, 2024.

[5] Teresa Scantamburlo, Atia Cortés, and Marie Schacht. Progressing towards responsible ai, 2020.

[6] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[7] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible ai, 2022.

[8] Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. Detectors for safe and reliable llms: Implementations, uses, and limitations, 2024.

[9] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation, 2023.

[10] John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus?, 2023.

[11] Si Cheng, Shuxu Huffman, Qingxiaoyang Zhu, Haotian Su, Raja Kushalnagar, and Qi Wang. "real learner data matters" exploring the design of llm-powered question generation for deaf and hard of hearing learners, 2024.

[12] David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. On classification with large language models in cultural analytics, 2024.

[13] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[14] Alejandro Tlaie. Exploring and steering the moral compass of large language models, 2024.

[15] Yi Zhang, Mengjia Wu, Guangquan Zhang, and Jie Lu. Responsible ai: Portraits with intelligent bibliometrics, 2024.

[16] Angelina Wang, Teresa Datta, and John P. Dickerson. Strategies for increasing corporate responsible ai prioritization, 2024.

[17] Abhishek Gupta. Making responsible ai the norm rather than the exception, 2021.

[18] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and Peter Grasch. Understanding alignment in multimodal llms: A comprehensive study, 2024.

[19] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models, 2023.

[20] James Prather, Paul Denny, Juho Leinonen, Brett A. Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, Stephen MacNeil, Andrew Peterson, Raymond Pettit, Brent N. Reeves, and Jaromir Savelka. The robots are here: Navigating the generative ai revolution in computing education, 2023.

[21] Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. Towards a responsible ai metrics catalogue: A collection of metrics for ai accountability, 2024.

[22] Necdet Gurkan and Jordan W. Suchow. Exploring public opinion on responsible ai through the lens of cultural consensus theory, 2024.

[23] Muneera Bano, Didar Zowghi, Pip Shea, and Georgina Ibarra. Investigating responsible ai for scientific research: An empirical study, 2023.

[24] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, and Michael Muller. A systematic literature review of human-centered, ethical, and responsible ai, 2023.

[25] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024.

[26] Wrick Talukdar and Anjanava Biswas. Improving large language model (llm) fidelity through context-aware grounding: A systematic approach to reliability and veracity, 2024.

[27] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation, 2024.

[28] Sara Court and Micha Elsner. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem, 2024.

[29] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices, 2021.

[30] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy, 2024.

[31] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge?, 2024.

[32] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[33] Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. Llm augmented llms: Expanding capabilities through composition. *arXiv preprint arXiv:2401.02412*, 2024.

[34] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

[35] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A review of multi-modal large language and vision models, 2024.

[36] Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier, and Ravit Dotan. Responsible ai for test equity and quality: The duolingo english test as a case study, 2024.

[37] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, and Zhenchang Xing. Towards a roadmap on software engineering for responsible ai, 2022.

[38] Eva Thelisson, Grzegorz Mika, Quentin Schneiter, Kirtan Padh, and Himanshu Verma. Toward responsible ai use: Considerations for sustainability impact assessment, 2023.

[39] Amna Batool, Didar Zowghi, and Muneera Bano. Responsible ai governance: A systematic literature review, 2023.

[40] Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. "a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms, 2024.

[41] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025.

[42] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said, 2024.

[43] Florian Scholten, Tobias R. Rebholz, and Mandy Hütter. Metacognitive myopia in large language models, 2024.

[44] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. The pursuit of fairness in artificial intelligence models: A survey, 2024.

[45] Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 1+1>2: Can large language models serve as cross-lingual knowledge aggregators?, 2024.

[46] Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Tanuja Ganu, Sunayana Sitaram, Kalika Bali, and Akshay Nambi. Bridging the gap: Dynamic learning strategies for improving multilingual performance in llms, 2024.

[47] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

[48] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[49] Yi Luo, Zhenghao Lin, Yuhao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. Ensuring safe and high-quality outputs: A guideline library approach for language models, 2024.

[50] Xinmeng Hou. Mitigating biases to embrace diversity: A comprehensive annotation benchmark for toxic language, 2024.

[51] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[52] Tomás Dodds, Astrid Vandendaele, Felix M. Simon, Natali Helberger, Valeria Resendez, and Wang Ngai Yeung. The impact of knowledge silos on responsible ai practices in journalism, 2024.

[53] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. Exploregen: Large language models for envisioning the uses and risks of ai technologies, 2024.

[54] Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and Azzurra Ragone. A rapid review of responsible ai frameworks: How to guide the development of ethical ai, 2023.

[55] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. Ai and ethics – operationalising responsible ai, 2021.

[56] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*, 2023.

[57] Melissa Ailem, Katerina Marazopoulou, Charlotte Siska, and James Bono. Examining the robustness of llm evaluation to the distributional assumptions of benchmarks, 2024.

[58] Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z. Pan. Trustscore: Reference-free evaluation of llm response trustworthiness, 2024.

[59] Muhammed Saeed, Elgizouli Mohamed, Mukhtar Mohamed, Shaina Raza, Muhammad Abdul-Mageed, and Shady Shehata. Desert camels and oil sheikhs: Arab-centric red teaming of frontier llms, 2024.

[60] Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. On the robustness of editing large language models, 2024.

[61] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. Are large language models reliable judges? a study on the factuality evaluation capabilities of llms, 2023.

[62] Betty Li Hou, Kejian Shi, Jason Phang, James Aung, Steven Adler, and Rosie Campbell. Large language models as misleading assistants in conversation, 2024.

[63] Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment, 2023.

[64] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer, 2019.

[65] Erick Galinkin. Towards a responsible ai development lifecycle: Lessons from information security, 2022.

[66] Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts, 2025.

[67] Julia Stoyanovich, Rodrigo Kreis de Paula, Armanda Lewis, and Chloe Zheng. Using case studies to teach responsible ai to industry practitioners, 2024.

[68] Shalaleh Rismani and AJung Moon. What does it mean to be a responsible ai practitioner: An ontology of roles and skills, 2023.

[69] Conrad Sanderson, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle. Towards implementing responsible ai, 2023.

[70] Séamus Lankford and Andy Way. Leveraging llms for mt in crisis scenarios: a blueprint for low-resource languages, 2024.

[71] Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, and Bela Gipp. Ai usage cards: Responsibly reporting ai-generated content, 2023.

[72] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[73] Nur Ahmed, Amit Das, Kirsten Martin, and Kawshik Banerjee. The narrow depth and breadth of corporate responsible ai research, 2024.

[74] Zhicheng Lin. Beyond principlism: Practical strategies for ethical ai use in research practices, 2024.

[75] Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models, 2024.

[76] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.

[77] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. How different groups prioritize ethical values for responsible ai, 2022.

[78] Stephanie Baker and Wei Xiang. Explainable ai is responsible ai: How explainability creates trustworthy and socially responsible artificial intelligence, 2023.

[79] Andrés Domínguez Hernández and Vassilis Galanos. A toolkit of dilemmas: Beyond debiasing and fairness formulas for responsible ai/ml, 2023.

[80] Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models, 2024.

[81] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[82] Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting, 2024.

[83] Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes, 2024.

[84] Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, Aiwei Liu, Yong Yang, Zhonghai Wu, Xuming Hu, Philip S. Yu, and Ying Li. A survey of aiops for failure management in the era of large language models, 2024.

[85] Milena Chadimová, Eduard Jurášek, and Tomáš Kliegr. Meaningless is better: hashing bias-inducing words in llm prompts improves performance in logical reasoning and statistical learning, 2024.

[86] Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems, 2024.

[87] Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. Llms4om: Matching ontologies with large language models, 2024.

[88] Shengsheng Qian, Zuyi Zhou, Dizhan Xue, Bing Wang, and Changsheng Xu. From linguistic giants to sensory maestros: A survey on cross-modal reasoning with large language models, 2024.

[89] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology, 2024.

[90] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

[91] Artūrs Kanepajs, Vladimir Ivanov, and Richard Moulange. Towards safe multilingual frontier ai, 2024.

[92] Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. Fairer preferences elicit improved human-aligned large language model judgments, 2024.

[93] Hayley Ross, Kathryn Davidson, and Najoung Kim. Is artificial intelligence still intelligence? llms generalize to novel adjective-noun pairs, but don't mimic the full human distribution, 2024.

[94] Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. Investigating automatic scoring and feedback using large language models, 2024.

[95] Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Wenhu Chen, and Ge Zhang. Chinese tiny llm: Pretraining a chinese-centric large language model, 2024.

[96] Ziyang Chen and Stylios Moscholios. Using prompts to guide large language models in imitating a real person's language style, 2024.

[97] Elisavet Koutsiana, Johanna Walker, Michelle Nwachukwu, Albert Meroño-Peñuela, and Elena Simperl. Knowledge prompting: How knowledge engineers use large language models, 2024.

[98] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, 2023.

[99] Divyanshu Kumar, Umang Jain, Sahil Agarwal, and Prashanth Harshangi. Investigating implicit bias in large language models: A large-scale study of over 50 llms, 2024.

[100] Pawel Robert Smolinski, Joseph Januszewicz, and Jacek Winiarski. Scaling technology acceptance analysis with large language model (llm) annotation systems, 2024.

[101] Euan D Lindsay, Mike Zhang, Aditya Johri, and Johannes Bjerva. The responsible development of automated student feedback with generative ai, 2025.

[102] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[103] Anka Reuel, Patrick Connolly, Kiana Jafari Meimandi, Shekhar Tewari, Jakub Wiatrak, Dikshita Venkatesh, and Mykel Kochenderfer. Responsible ai in the global context: Maturity model and survey, 2024.

[104] Petar Radanliev and Omar Santos. Ethics and responsible ai deployment, 2023.

[105] Conrad Sanderson, David Douglas, and Qinghua Lu. Implementing responsible ai: Tensions and trade-offs between ethics aspects, 2024.

[106] Farouq Sammour, Jia Xu, Xi Wang, Mo Hu, and Zhenyu Zhang. Responsible ai in construction safety: Systematic evaluation of large language models and prompt engineering, 2024.

[107] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[108] Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge barriers in multilingual large language models, 2024.

[109] Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. Few-shot fairness: Unveiling llm's potential for fairness-aware classification, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.