# A Survey of LLM Accountability in Education, Healthcare, and Human Resources

**Abdolrahim Arjomand, Oscar Granadino Horvath, Rui Sun, Weihan Qu**
University of California, Los Angeles
abdolrahimarj@g.ucla.edu, oscargh@g.ucla.edu, ruis@g.ucla.edu, weihanqu@g.ucla.edu

## Abstract

The rapid deployment of Large Language Models within different sectors has seen tremendous growth. However, the scale at which they are being adopted raises concerns about stakeholder accountability and governance. In particular, high-stakes sectors such as the education, healthcare, and human resources sectors have LLMs impacting human life directly. Some governance framework attempts have been introduced such as the SB1047 act, the EU AI act, and the AB-2013 bill but they are still in the early stages from being considered unified and ready. We believe that, as a first step, a comprehensive survey is required to further explore these three sectors to offer researchers valuable literature references to build upon. In this survey, we provide an overview of how LLM accountability is currently managed in the education, healthcare, and human resources sectors, key findings, gaps, and our insights and conclusions about what should be done further.

## 1 Introduction and Motivations

The proliferation of Large Language Models (LLMs) in the last few years has been profound enough to gather the attention of many users and stakeholders across many different sectors (Urlana et al., 2024). LLMs have redefined the process of information generation and utilization at unprecedented levels. Important sectors such as the education sector, healthcare sector, and the labor economy sector have had LLMs reshaping some of their fundamental processes whether it be in personalization of education, decision making in clinical care, or driving hiring algorithms. However, the rapid development and use of LLMs highlights critical concerns about stakeholder accountability, governance, and regulatory oversight (Ferdaus et al., 2024). This is especially true in high stakes applications where decisions impact human lives directly. In the education sector the impacted stakeholders are students, teachers, and model developers. In the healthcare sector, the main impacted stakeholders include but are not limited to patients, medical practitioners, and model developers. When it comes to the human resources sector, hiring algorithms impact people of different genders, races, religious, and ethnic backgrounds, especially in the areas of privacy and fairness.

There are some accountability frameworks which have been proposed such as the SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act for the state of California (Wiener et al., 2024). Although this act failed to be passed into law, we still think it is necessary to showcase what it says about the critical requirements for model developers to adhere to. Some of these requirements include responsible, harm-free development of AI systems for the general public, introduction of auditing cycles, and also the creation of the Government Corporations Agency consortium responsible for enhancing the oversight of AI innovation in a safe and ethical manner (Wiener et al., 2024). Furthermore, we find some similarities between the EU AI Act and the AB-2013 Generative Artificial Intelligence: Training Data Transparency bill (European Commission, 2024; California State Legislature, 2024). Both regulatory frameworks, although from geographically distant locations, emphasize the importance of data transparency to the user. This also includes making it clear to the user what data has been used to train the models they are using. The EU AI act also emphasizes that any detrimental effect caused by AI models falls under the full responsibility of model developers (European Commission, 2024).

Although some accountability frameworks have been proposed there is still a lack of a comprehensive survey that shows how different sectors are managing accountability and tackling LLM based risks. Our goal is to create a survey paper showcasing how the current education, healthcare, and human resources sectors incorporate AI-LLM accountability while highlighting gaps and suggest-

ing areas of improvement. We would also like to provide valuable literature references and analytical viewpoints for researchers in this field while calling for greater attention to the importance of LLM accountability. The outline of our survey paper is structured as Figure. 1 in the appendix: LLM accountability in education, LLM accountability in healthcare, and LLM accountability in human resources.

## 2 LLM Accountability in Education

### 2.1 Overview

Large Language Models (LLMs) are increasingly being integrated into educational tools, and this trend offers new opportunities but also raises critical issues of accountability. This section examines recent research and policy reports on two key facets of LLM accountability in education: (1) fairness in automated grading and (2) ethical concerns in AI-assisted learning.

### 2.2 Fairness in Automated Grading with LLMs

Automated grading systems (like essay scorers) promise consistency and efficiency, but studies have found they can reproduce or even amplify human biases (Chinta et al., 2024). Early research on automated essay scoring (AES) and related tools (even before LLMs) revealed systematic bias: for example, algorithms sometimes over-score or under-score student work based on demographics such as gender, race, or socioeconomic status (Schaller et al., 2024). Schaller et al. (2024) showed that both traditional machine-learning and newer deep-learning AES models gave unfairly higher or lower scores to certain groups of K-12 students. These findings underscore that algorithmic bias in grading is a real concern. To promote accountability, researchers and practitioners are proposing solutions to make AI grading fairer. A common theme is that fairness should be treated as a first-class objective in model development, not an afterthought. For example, Fenu et al. (2022) and others argue that LLM-based tools should be trained on data that represent all student groups and should be evaluated with fairness metrics alongside accuracy (Schaller et al., 2024; Wang et al., 2024).

### 2.3 Ethical Concerns in AI-Assisted Learning

Beyond grading, LLM-powered systems are being used as tutors and student support tools. This raises several ethical and accountability concerns – notably around student privacy and transparency of AI decisions. Recent research and policy reports highlight these issues and offer guidance on responsible AI use in schools.

**Student Privacy and Data Protection.** AI-driven learning platforms often rely on extensive student data (e.g. learning behaviors, performance history) to personalize instruction. This brings significant privacy risks. Huang (2023) warns that the rapid adoption of AI in education has outpaced data protection measures, leading to concerns about unauthorized access or misuse of sensitive student information. Privacy scholars emphasize that students have the right to control their personal data, yet many AI tools operate as "black boxes" where it's unclear what data is collected or how long it is retained. This is especially problematic given that educational data can include personal identifiers and even infer sensitive attributes.

**Transparency and Explainability in AI Decisions.** Due to the "black box" nature of LLM, Milano et al. (2023) notes that opacity in LLM-driven systems makes it hard to detect errors or biases and erodes user trust (Wang et al., 2024). Thus, we need to improve transparency. Recent work in explainable AI (XAI) is beginning to intersect with educational applications. Some studies (Finlayson et al., 2024) propose techniques to probe LLMs and reveal aspects of their internal reasoning or decision criteria. For example, researchers have experimented with prompting an LLM to explain its grade in natural language, or using smaller interpretable models alongside the LLM to audit its decisions. While perfect explainability for large neural networks remains an open challenge, even partial transparency measures (like providing feature importances or example-based explanations) can help.

## 3 LLM Accountability in Healthcare

### 3.1 Overview

Large Language Models (LLMs) are rapidly reshaping healthcare through applications ranging from diagnostic decision support to automated medical report generation. While these models offer significant promise in enhancing patient care and operational efficiency, their integration also raises serious accountability concerns. In this section, we survey three key accountability areas in healthcare: bias in medical diagnosis, patient data privacy, and fairness in healthcare applications.

### 3.2 Bias in Medical Diagnosis

LLMs applied to diagnostic tasks can inadvertently inherit and even amplify biases present in their

training data. For instance, several studies have observed that models trained on imbalanced medical image datasets tend to under diagnose conditions in marginalized groups—such as female patients or certain ethnic minorities—due to their underrepresentation in the training corpus (Singhal et al., 2023). Such bias not only undermines the reliability of automated diagnostics but may also perpetuate existing health disparities. To mitigate these risks, researchers advocate that fairness must be treated as a first-class objective during model development. Techniques such as importance weighting, domain adaptation, and fairness-aware training have been proposed to rebalance data representation and adjust decision thresholds appropriately (Zhou et al., 2023a,b). Although these strategies show promise, the inherent complexity of medical data and the challenge of accessing diverse, high-quality datasets mean that completely eradicating bias remains an ongoing concern (Gao et al., 2023).

### 3.3 Patient Data Privacy

The success of LLMs in healthcare is heavily predicated on the availability of rich clinical datasets, including Electronic Health Records (EHRs) and clinical notes. However, these datasets are highly sensitive, containing personal and medical details that demand stringent privacy safeguards. Recent findings suggest that LLMs can inadvertently memorize and later reproduce fragments of their training data, posing significant risks of exposing patient information (Ouyang et al., 2022). In response, several privacy-preserving techniques have been introduced. For example, differential privacy—by adding calibrated noise during training—helps ensure that individual patient records remain indistinguishable from the aggregate data (Mahajan et al., 2020). Additionally, federated learning frameworks enable decentralized training on local datasets without transferring raw data to central servers, thereby further reducing the risk of data leakage (Guo et al., 2022). While these methods mark substantial progress, they often involve trade-offs between model performance and privacy guarantees, underscoring the need for continuous evaluation and improvement of privacy measures in clinical AI systems (Ouyang et al., 2022).

### 3.4 Fairness in Healthcare Applications

Beyond diagnostics, LLMs are increasingly employed in broader healthcare applications, including patient triage, treatment recommendation, and resource allocation. Ensuring fairness across these applications is critical to prevent the reinforcement of existing inequities in healthcare delivery. Studies have shown that without deliberate interventions, AI systems may systematically favor certain demographic groups over others—leading to unequal access to care and suboptimal treatment recommendations for underrepresented populations (Stade et al., 2024). To promote fairness, it is essential to incorporate regular algorithmic audits, transparent evaluation metrics, and inclusive training datasets that reflect the diversity of patient populations. Furthermore, interdisciplinary collaborations among clinicians, data scientists, and ethicists are vital to developing governance frameworks that enforce fairness at every stage of AI development and deployment. Such collaborative efforts help ensure that AI tools not only meet technical benchmarks but also align with societal values of equity and justice in healthcare (Stade et al., 2024; Omiye et al., 2023).

## 4 LLM Accountability in Human Resources

### 4.1 Overview

Artificial intelligence tools have recently become essential in various Human Resources (HR) activities, such as hiring, performance reviews, and employee development. While these systems can help organizations work more efficiently, they also pose serious questions about fairness and accountability. This section highlights current findings on two main areas related to AI accountability in HR: (1) ensuring fair decisions in recruitment and (2) recognizing ethical risks in AI-based employee support.

### 4.2 Fair Decision-Making in Recruitment

AI-powered recruitment platforms promise quicker and more consistent hiring processes, yet research shows these tools can replicate or even exacerbate human biases (Bogen and Rieke, 2018). For example, Tsamados et al. (2021) describe numerous cases where algorithmic job ads in the tech industry were shown more frequently to men than to women. One proposed solution is to ensure that datasets are gathered in a deliberate, balanced way (Hanson et al., 2023), so that overrepresented groups do not skew the training data.

### 4.3 Ethical Concerns in LLM-Driven HR Support

Beyond hiring, AI is being used in areas like performance management and personalized employee

development. Although these applications can simplify HR processes, they also raise ethical issues regarding privacy and transparency. Recent studies emphasize the difficulty of protecting sensitive employee information when AI systems collect large amounts of performance data (Tambe et al., 2019). In many cases, employees have limited insight into how personal metrics are gathered or used, creating possible legal and reputational risks if that data is ever misused.

**Privacy and Data Handling.** LLM-based HR tools generally depend on extensive employee data such as productivity indicators, communication logs, and training progress. While the ability to track these metrics allows for more tailored support, it also poses serious privacy threats. Ajunwa et al. (2017) argue that workforce analytics have outpaced existing privacy regulations, calling for more stringent data management policies. Moreover, some AI solutions remain proprietary and opaque, leaving employees in the dark about what exactly is collected and how (Chen et al., 2023). Recently, there has been a growing trend of using sentiment analysis tools to track how employees interact with clients, offering valuable insights into customer-facing conversations Gelbard et al. (2018). While these tools can help organizations evaluate performance and enhance service quality, they also raise significant ethical and privacy concerns that require careful consideration.

**Transparency and Explainability in LLM Decisions.** Many HR-oriented AI systems operate as "black boxes," making it difficult to spot biases or errors. Leicht-Deobald et al. (2019) recommend using more interpretable or transparent algorithms to detect unfair patterns in decision-making processes. Although achieving complete transparency in complex LLMs remains challenging, even partial explanations -such as highlighting key factors in a performance score- can improve trust and let human supervisors intervene when necessary (Raghavan et al., 2020).

**Frameworks.** Recent legislative measures, including the EU AI Act (European Commission, 2021) and California's AB-2013 (California State Legislature, 2023a), underscore the need for transparent datasets and responsible data usage when deploying AI in HR. Under the EU AI Act, recruitment systems leveraging AI are deemed high-risk, thus requiring strict data reporting, human oversight, and candidate notifications whenever AI tools are used. The law also grants applicants the right to question or challenge AI-driven hiring decisions if they suspect errors.

Although the proposed SB 1047 (California State Legislature, 2023b) did not pass due to political and feasibility concerns, it laid out a framework featuring mandatory bias checks, transparent AI decision reporting, and human oversight in the hiring process. While this measure was rejected, its ideas have influenced ongoing discussions aimed at preserving fairness, privacy, and trust in AI-powered HR systems.

## 5 Conclusion

LLMs are being rapidly deployed in many sectors and are also becoming a fundamental part of many day to day processes. However, they are not without limitations. This paper focused on providing a comprehensive survey highlighting the impact of LLM accountability in education, healthcare, and human resources.

**LLM Accountability in Education.** LLMs hold great promise for enhancing education – from streamlining grading to powering intelligent tutors – but realizing this promise requires confronting issues of fairness, ethics, and transparency. The recent literature makes clear that fairness in automated grading is an urgent concern: without deliberate safeguards, LLM-based graders can perpetuate biases, undermining the legitimacy of assessments. Likewise, student privacy must be rigorously protected in AI-enabled learning environments; trust can quickly evaporate if learners feel surveilled or at risk of data abuse. The past three years have seen not only warnings about these challenges but also concrete steps toward solutions – from technical bias mitigation techniques to ethical guidelines issued by governments and organizations. A recurring theme is the call for transparency and human oversight at every stage: if educators understand and can intervene in AI processes, accountability is vastly improved. Continued collaboration between AI researchers, educators, and policymakers will be vital to ensure that LLMs serve as tools for educational advancement without compromising fairness, privacy, or informed, transparent decision-making.

**LLM Accountability in Healthcare.** LLMs hold transformative potential for enhancing healthcare delivery, yet their benefits can only be fully realized if accountability challenges are addressed head-on. Mitigating bias in medical diagnosis, safeguarding patient data privacy, and embedding fair-

ness into healthcare applications are essential pillars for responsible AI deployment. In our view, a multidisciplinary approach—combining technical innovation with ethical oversight and rigorous regulatory compliance—is critical. By continually refining these accountability frameworks and engaging a broad range of stakeholders, the healthcare community can ensure that LLMs contribute positively to patient outcomes while upholding the highest standards of equity and trust.

**LLM Accountability in Human Resources.** LLMs offer major benefits for reshaping HR departments, speeding up recruitment processes and supporting employee development. However, this potential also comes with serious responsibilities. Recent findings show that bias can unintentionally be embedded in AI hiring tools, emphasizing the need for continuous monitoring to prevent discrimination. Furthermore, handling employee data securely remains critical, and strong privacy protections are essential. Fortunately, there is growing awareness of these challenges, and various solutions have appeared. These include methods to reduce bias and industry guidelines that call for greater transparency in AI-driven processes. One key point in the research is that human oversight is necessary at every stage of AI use. By giving HR staff the knowledge to understand how these systems function, and letting them step in when needed, it becomes much easier to keep these technologies accountable. As AI developers, HR leaders, and policymakers work together, AI is more likely to boost HR practices while still preserving fairness, privacy, and reasoned human judgment.

## Embedded Ethics Discussion

Addressing our topic involves teaching students how accountability frameworks apply in education, healthcare, and human resources sectors. To further explore how we can convey our message to people who are starting to learn about this topic, we propose a set of structured lectures to enhance theoretical understanding followed by hands-on coding assignments. The lectures would be outlined as follows: introduce LLM accountability, present case studies of LLM accountability in education, healthcare, and human resources, offer students reading assignments while highlighting the shortcomings of LLM accountability in these respective sectors, then introduce current LLM accountability frameworks and the gaps that exist within them, and finally end with a final project where students research and either propose new or make amendments to these frameworks. Along the first half of the quarter a coding assignment would be assigned with the following content: offering a skewed dataset that is biased for an HR hiring algorithm and asking students to implement bias reduction techniques with the aim of meeting a statistical equality metric where no group of datasets are favored over others.

## Contribution Statement

Abdolrahim Arjomand: Responsible for the abstract, introduction and motivations, and embedded ethics section (literature review, presentation, paper writing). Provided the acts and bills of established accountability frameworks. Contributed to the conclusion section.

Oscar Granadino Horvath: Responsible for LLM accountability in the HR section (including the literature review, presentation, and paper writing), peer reviewing the introduction section, and contributing to the conclusion section.

Rui Sun: Propose the framework of the survey paper; Responsible for LLM Accountability in Education section (literature review, presentation, paper writing) and Conclusion section.

Weihan Qu: Responsible for LLM accountability in Healthcare section (literature review, presentation, paper writing). And contributed to the conclusion section.

We believe that everyone has a reasonable contribution to the course project.

## Acknowledgements

We appreciate the feedback from the professor and classmates, and we feel fortunate to have such a great team.

## References

Ifeoma Ajunwa, Kate Crawford, and Jason Schultz. 2017. Limitless worker surveillance. *California Law Review*, 105:735–776.

Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. Upturn.

California State Legislature. 2023a. AB-2013 Generative Artificial Intelligence: Training Data Transparency. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013.

California State Legislature. 2023b. SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence

Models Act. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047.

California State Legislature. 2024. Ab-2013 generative artificial intelligence: training data transparency. Assembly Bill No. 2013.

Jiawei Chen, Shiyu Wang, and Timothy White. 2023. Human-centered design to address biases in ai. *Journal of Medical Internet Research*, 25:e42991.

Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, Tai Le Quy, and Wenbin Zhang. 2024. Fairaied: Navigating fairness, bias, and ethics in educational ai applications. *arXiv preprint arXiv:2407.18745*.

European Commission. 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

European Commission. 2024. Article 50: Transparency obligations for providers and deployers of certain ai systems. *EU Artificial Intelligence Act*.

Gianni Fenu, Roberta Galici, and Mirko Marras. 2022. Experts' view on challenges and needs for fairness in artificial intelligence for education. In *International Conference on Artificial Intelligence in Education*, pages 243–255. Springer.

Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Ioup, Kendall N Niles, Ken Pathak, and Steven Sloan. 2024. Towards trustworthy ai: A review of ethical and robust large language models. *arXiv preprint arXiv:2407.13934*.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. 2024. Logits of api-protected llms leak proprietary information. *arXiv preprint arXiv:2403.09539*.

Weihao Gao et al. 2023. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*.

Roy Gelbard, Roni Ramon-Gonen, Abraham Carmeli, Ran M. Bittmann, and Roman Talyansky. 2018. Sentiment analysis in organizational work: Towards an ontology of people analytics. https://doi.org/10.1111/exsy.12289. First published: 25 May 2018.

Quan Guo et al. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564.

Vicki L. Hanson, Rosangela Bittencourt, and Juliana Freire. 2023. Garbage in, garbage out: Mitigating risks and maximizing benefits of ai in research. *Nature*, 620:29–32.

Lan Huang. 2023. Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers*, 16(2):2577–2587.

Ulrich Leicht-Deobald, Timo Busch, Christoph Schank, Andreas Weibel, Susanne Schafheitle, Isabelle Wildhaber, and Gerhard Kasper. 2019. The challenges of algorithm-based hr decision-making for personal integrity. *Journal of Business Ethics*, 160(2):377–392.

Diwakar Mahajan et al. 2020. Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning. *JMIR Medical Informatics*, 8(11):e22508.

Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334.

Jesutofunmi A. Omiye et al. 2023. Large language models in medicine: the potentials and pitfalls.

Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221.

Karan Singhal et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Elizabeth C. Stade et al. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42.

Antonio Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. 2021. The ethics of algorithms: Key problems and solutions. *AI & Society*, 36:1–17.

Ashok Urlana, Charaka Vinayak Kumar, Ajeet Kumar Singh, Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala, and Rahul Mishra. 2024. Llms with industrial lens: Deciphering the challenges and prospects–a survey. *arXiv preprint arXiv:2402.14558*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Scott Wiener, Richard Roth, Susan Rubio, and Henry Stern. 2024. Sb 1047: Safe and secure innovation for frontier artificial intelligence models act.

Hongjian Zhou et al. 2023a. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Juexiao Zhou et al. 2023b. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *arXiv preprint arXiv:2306.10765*.

# 6 Appendix

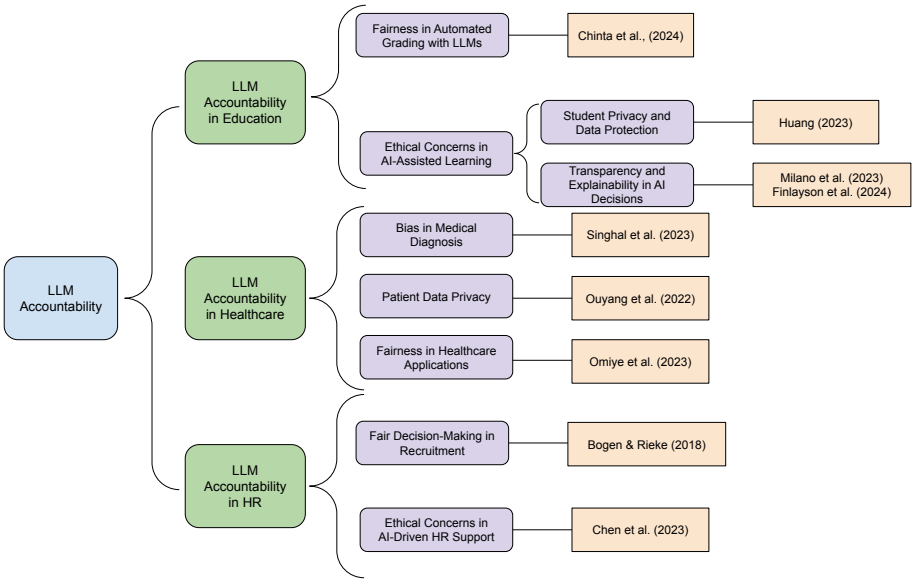


Figure 1: A taxonomy of LLM accountability in education, healthcare, and human resources with representative works.