

Tensions Between Biomedical Ethics Review and Responsible AI Research at University Hospitals in South America

Sofia Martinelli^{1,2,3,*}, Luciana Benotti^{1,2,3,6}, Viviana Cotik^{3,5} and Lorena Etcheverry^{4,6}

¹Universidad Nacional de Córdoba, Argentina

²Fundación Vía Libre, Argentina

³CONICET, Argentina

⁴Instituto de Computación, Universidad de la República, Uruguay

⁵Universidad de Buenos Aires, FCEN, Departamento de Computación, Argentina

⁶CICADA, Universidad de la República, Uruguay

Abstract

This paper examines the alignment between existing biomedical ethics review processes and Responsible AI research involving Large Language Models (LLMs) in healthcare. To explore this subject, we analyze the ethics reviews of two related projects in two South American countries. The first case study, on ICD-10 coding of clinical records, is being developed within the Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático (CICADA, Universidad de la República) in collaboration with the Hospital de Clínicas in Montevideo, Uruguay, and was reviewed by the Ethics Committee of that hospital. The second case study, on simulated virtual patients for medical education, is being developed by the research group Contextualized AI at the Universidad Nacional de Córdoba together with the Hospital Nacional de Clínicas in Córdoba, Argentina, and was reviewed by the Institutional Ethics Board of the Universidad Nacional de Córdoba. The two cases differ in an informative way: the first processes real patient records, while the second relies on fictional clinical scenarios. In both cases, the review processes revealed methodological and conceptual tensions between traditional biomedical ethics review and contemporary Responsible AI research. In particular, both reviews focused on protecting research participants and data subjects, while several risks central to Responsible AI fall on people beyond these groups, such as the social groups represented in model outputs and the future patients of medical students.

Keywords

Biomedical Ethics, Responsible AI, Large Language Models, Healthcare Governance

1. Introduction

Large Language Models (LLMs) are progressively investigated for healthcare applications, including summarizing electronic health records, generating reports, and answering questions over clinical data. Their use in healthcare, however, raises significant concerns regarding hallucinations, bias, explainability, opacity, privacy, and accountability. These issues have become central to discussions on Responsible AI in medicine [1, 2, 3, 4]. Consequently, research involving LLMs in clinical contexts is now frequently framed within Responsible AI principles.

This paper reports on two related projects that share some of their members and a common approach to Responsible AI. The first project is being developed within the Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático (CICADA, Universidad de la República) in collaboration with the Hospital de Clínicas in Montevideo, Uruguay. Among other objectives, it aims to identify clinical concepts that can be mapped to ICD-10¹ codes in anonymized clinical records. The project brings together researchers from computer science, medicine, and the social sciences to address questions

First Workshop on Ethics in Clinical AI Applications (ETHICAIA), co-located with the 35th International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026), Bremen, Germany

*Corresponding author.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹ICD-10, the International Classification of Diseases, Tenth Revision, is the classification system maintained by the World Health Organization (WHO) for coding diseases, health conditions, and causes of death. It is used for both morbidity and mortality statistics. <https://icd.who.int/browse10/2019/en>

about fairness, inclusion, and gender bias in AI systems. It draws on the Responsible AI considerations discussed in [5, 6]. In this context, Responsible AI refers to the alignment of algorithmic performance with fundamental human values and safety standards. The second project is being developed by the research group Contextualized AI at the Universidad Nacional de Córdoba (UNC), together with the Hospital Nacional de Clínicas in Córdoba, Argentina. It evaluates the use of LLMs to simulate patients for medical education. The CICADA project also plans to explore patient simulation, but this component has not yet been submitted to the Ethics Committee of the Hospital de Clínicas.

Each project underwent its own ethical review. The CICADA project was reviewed by the Ethics Committee of the Hospital de Clínicas and was officially registered with Uruguay’s Ministry of Public Health (MSP). The review required detailed specifications concerning datasets, participant involvement, anonymization procedures, data storage, institutional responsibilities, and evaluation workflows. The project on simulated patients received formal approval from the Institutional Health Research Ethics Committee of the Hospital Nacional de Clínicas (CIEIS HNC), which established requirements concerning voluntary participation, informed consent, and the deletion of interaction data. These requirements were legitimate and consistent with established biomedical ethics principles related to confidentiality and consent.

Comparing the two reviews is informative because the cases differ in a central aspect: the first processes real clinical records, while the second relies on fictional clinical scenarios. Despite this difference, both reviews concentrated on the people directly involved in the research, while several risks central to Responsible AI remained outside their scope. We will move into the tensions shortly but first we briefly describe the projects and their use cases.

2. Research Projects and Use Cases

The two use cases described below both involve LLMs in healthcare, but they raise distinct ethical considerations. For each case we describe the project, the requirements established by the corresponding ethics review, and the Responsible AI risks that we identified but that the review did not explicitly address.

2.1. Case 1: ICD-10 Coding at the Hospital de Clínicas, Montevideo

The CICADA project investigates architectures that combine local deployment of open-source LLMs, structured clinical knowledge such as SNOMED CT² and knowledge graphs, and privacy-preserving mechanisms to reduce unnecessary exposure of sensitive clinical data. A central objective is not only to develop AI-based clinical tools but also to identify which healthcare tasks are genuinely suitable for LLM-based assistance and under what technical and ethical conditions. Rather than presuming widespread adoption of generative AI, the project adopts a minimization principle: large-scale language models should be employed only when they provide substantial advantages over simpler, more interpretable alternatives. This approach aligns with recent critiques that argue high-level AI principles are insufficient without concrete governance and risk-assessment mechanisms [7].

ICD-10 codes are commonly used to represent the diagnoses described in clinical texts in a standardized form, which enables their statistical analysis [8, 9]. This use case has two related goals: to evaluate the quality of the ICD-10 codes already assigned to clinical records, and to explore the potential use of large and small language models (LLMs and SLMs) to automatically assign ICD-10 codes to clinical notes. The two goals are connected, since evaluating automatically generated codes requires a model of what a good code assignment is. Most work on automated coding evaluates systems with standard measures such as precision, recall, and F1, which treat all errors equally. However, ICD-10 is a taxonomy, and not all errors carry the same weight: assigning a code from the same category as the correct one is different from assigning a code from an unrelated chapter. Some evaluations already

²SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is a comprehensive clinical terminology with ontological structure, comprising concepts, descriptions, and semantic relationships among them. <https://www.snomed.org/>

account for this, either by computing scores at the category level [10] or through hierarchical measures [11, 12], and recent work weights errors by their financial impact [13]. Distance in the taxonomy and financial cost, however, do not necessarily reflect the consequences of an error for clinical care or health statistics, so deciding how errors should be weighted is a value judgment that requires the participation of physicians and clinical coders. The quality of existing codes is also a concern: codes commonly treated as gold standards contain substantial undercoding and errors [14, 15], and recent work has begun to audit them automatically, including with LLMs [15, 16]. Finally, the main benchmark for ICD-10 coding in Spanish is built from published clinical case reports rather than routine hospital records, and its organizers point to the lack of coded clinical records from hospitals [10]. Our project aims to address these gaps by working with routine clinical records from the Hospital de Clínicas and by defining, together with physicians, how coding errors should be weighted. In line with the minimization principle, comparing LLMs with smaller models allows us to assess whether the additional cost and opacity of larger models are justified for this task. To evaluate model performance, a gold standard will be required. If the quality of the existing manually-assigned codes is insufficient, this gold standard will be developed in collaboration with physicians, first defining annotation criteria and then annotating clinical records accordingly. Since the same models may be involved both in assessing existing codes and in generating new ones, a subset of records annotated by human experts, with measured agreement between annotators, will serve as a reference for both tasks and help avoid circular evaluation.

The use of these models for this task could help reduce the workload associated with manually assigning codes and provide tools to improve the quality of data in health information systems. However, implementing them requires careful examination of personal data protection, potential biases introduced by the models, and fairness, including an assessment of their differential impact across sex and age groups.

The ethics committee requested the implementation of standard measures for the protection of clinical records, including data anonymization and the storage and processing of data within the hospital's infrastructure. These requirements are fundamental and represent a first step towards data protection and methodological rigor. Nevertheless, Responsible AI considerations were not explicitly addressed. All project participants will be required to sign a confidentiality agreement. Additionally, we propose de-identifying clinical records before granting access to project members, applying different procedures to structured and unstructured data. Finally, all LLMs and SLMs will be run on local infrastructure rather than in the cloud. In addition to meeting these requirements, our work will specifically examine potential biases and hallucinations, assessing model fairness, and evaluating the performance of LLMs and SLMs for this task. All these aspects are fundamental to the responsible use of AI.

2.2. Case 2: Simulated Virtual Patients for Medical Education at UNC

The second use case, developed at UNC together with the Hospital Nacional de Clínicas, explores the use of LLM-based simulated virtual patients (SVPs) to support medical education. Clinical anamnesis is a fundamental skill in medical training, traditionally practiced through interactions with real or simulated patients. However, these approaches face limitations related to operational costs, availability of trained actors, and difficulties in scalability [17, 18, 19]. LLM-based conversational agents can act as SVPs capable of sustaining interactions through natural language, potentially combining the flexibility of simulated patients with the reproducibility of standardized patients [20, 21, 22]. This proposed application raises ethical considerations that differ from those associated with processing real clinical records.

Beyond overcoming logistical barriers, SVPs directly alleviate a central ethical conflict in medical education: the necessity of using real patients as training subjects for inexperienced learners. Practicing anamnesis on actual patients inevitably introduces risks of patient discomfort, cognitive burden, or substandard communication, creating a tension between student learning needs and patient welfare. By shifting early-stage practice to an LLM-driven environment, SVPs offer a practice setting that does not involve real patients. Here, students can develop clinical empathy, experiment with diverse communication strategies, and learn from mistakes without exposing real, often vulnerable, patients to

the learning curve of novice clinicians. However, this setting introduces risks of its own.

Unlike the first use case, this approach does not rely on real patient records. Instead, it uses fictional clinical scenarios developed by physicians drawing on their clinical and teaching experience. These scenarios are based on cases commonly used in medical education, adapted to reflect diverse sociocultural contexts across Argentina and to create socially and culturally situated patient profiles. This approach aims to provide realistic clinical simulations while protecting patient privacy. Although the simulations would not involve real patients, LLM-generated responses may introduce other risks. For instance, LLMs may reproduce stereotypes or inaccurately represent patients from different demographic and sociocultural backgrounds, particularly in contexts such as South America [23]. Specifically, LLMs may exhibit an excess of negative politeness [24], such as over-emphasizing non-imposition, which deviates from local clinical pragmatics in Argentina, where patients typically demonstrate higher directness during history-taking (see Figure 1). As noted by a participating clinical professor, real patients routinely hand over medical records directly rather than asking permission to share them. Such misalignments also raise a risk for learning itself: students may adapt their communication to the pragmatics of the simulated patient rather than to those of real patients, which could later affect their interactions in clinical practice.

Dialogue Excerpt: Excessive Negative Politeness in Simulated Patient Interaction

Healthcare Professional: Have you had any check-ups or medical tests done?

Simulated Patient: Yes, let me see... my daughter left me some things here...
I have recent test results, *in case you need them.*

Expert Educator Feedback (Medical Faculty Professor):

"Typically, the patient hands over their test results directly without asking whether I want to see them."

Figure 1: Excerpt illustrating an excess of negative politeness (non-imposition: *"in case you need them"*) generated by the SVP. Note: The original interaction occurred in Spanish (*"Tengo estudios recientes, por si los necesita"*); this English translation illustrates the pragmatic misalignment flagged by a clinical professor as non-idiomatic for the local healthcare context in Argentina.

Beyond stereotypes and cultural inappropriateness, LLM-based virtual patients may introduce other ethical risks. First, environmental sustainability concerns arise from the compute requirements of large language models; we mitigate this carbon footprint and energy expenditure by deactivating extended reasoning modes during inference and investigating conversational realism under constrained computation. Second, although medical faculty and students are trained in data protection, the deployment of LLM interfaces carries a latent risk of unintentional disclosure of protected health information. To address this, our framework relies on open-weights models hosted entirely on local, secure university infrastructure, ensuring that data remains within institutional control. Third, deploying institutional LLMs creates financial uncertainty regarding infrastructure maintenance and computational overhead. Fourth, there is a risk of forcing new technology onto students, which the requirements of the ethics review described below address directly.

As an initial step, an exploratory activity was conducted with physicians involved in medical education to investigate the potential of LLM-based simulated virtual patients, without directly involving students. This pilot evaluation is described in [25]. This approach aimed to minimize potential ethical risks by exploring the use of LLM-based simulations with physicians before involving students.

The study with students received formal ethical approval from the CIEIS, which established the following requirements. Participation is strictly voluntary and requires informed consent. Participants are provided with an informed consent form explaining the purpose of the research and how their interaction data will be collected, anonymized, and used for research purposes. Non-participation carries no academic penalty or impact on course evaluation, and participants may withdraw at any

point, triggering the complete deletion of their interaction data. Moreover, since the educational use of SVPs remains under investigation, the activity is intended as a voluntary learning opportunity rather than a formal assessment tool, and performance during the simulations will not be used for academic assessment. As in the first use case, these requirements focus on the protection of research participants. The review did not explicitly address the risks that originate in the model outputs, such as stereotyped or culturally inappropriate representations of patients.

3. Discussion: Challenges and Tensions of Ethical Review

The ethical review processes of both projects revealed two main tensions discussed below.

The first tension was a methodological mismatch between traditional biomedical ethics review and exploratory Responsible AI research. Existing review frameworks typically assume stable protocols, predefined workflows, and fully specified interventions. In contrast, Responsible AI research involving LLMs is often iterative and adaptive, requiring ongoing evaluation and governance processes [26, 5]. The CICADA project aims to investigate appropriate use cases, strategies for mitigating hallucinations, safe architectures, and methods for evaluating fairness and explainability. Many of the methodological and governance mechanisms requested during ethics review are themselves subjects of ongoing research and development. For instance, the request for a justified sample size for ICD-10 coding is important for methodological rigor, but defining such thresholds in advance is difficult when one of the goals of the project is precisely to determine whether, and under which conditions, LLMs are suitable for the task. Similarly, in the second case, the simulated patients are refined iteratively in response to feedback from educators, as in the example of Figure 1, so the system used at the end of the study may differ from the one initially reviewed.

Second, the review processes exposed a conceptual gap between classical bioethics and Responsible AI governance. While the committees emphasized privacy, anonymization, confidentiality, and consent, they did not explicitly address several risks central to Responsible AI, such as hallucinations, algorithmic bias, lack of explainability, uncertainty communication, or governance of model outputs [2]. The comparison between our two cases makes one aspect of this gap especially visible. Biomedical ethics review is centered on research participants and data subjects: patients whose records are processed in the first case, and students who interact with the system in the second. However, several harms of LLMs fall on people who are neither. In the second case, no real patient data is processed and students are protected through consent and withdrawal, yet the sociocultural groups portrayed by the simulated patients may be represented through stereotypes, and the future patients of students trained with these systems may be affected by what students learn. In the first case, biased coding may affect the quality of health statistics, and with it the groups that are poorly represented in them, beyond the individuals whose records are processed. These risks lie outside the scope of the reviews as currently designed. We do not see this as a failure of the committees, whose mandate and expertise are defined around the protection of participants, but as a structural gap. Recent studies indicate that artificial intelligence systems in healthcare require broader, more continuous governance models that go beyond traditional bioethical review, focusing on lifecycle-based oversight and distributed accountability frameworks [27].

This experience does not indicate that biomedical ethics frameworks are inadequate, but rather that they may be incomplete for evaluating health research involving LLMs. In both of our projects, the research teams themselves currently design and carry out assessments of bias, hallucinations, and cultural appropriateness, without external oversight. Our case studies suggest that future approaches to the governance of generative AI in healthcare may need to combine traditional bioethical principles with iterative mechanisms for assessing responsible AI, such as phased approvals, continuous oversight, experimentation in test environments, and explicit assessments of algorithmic harms and biases, including harms to people who are not research participants.

These tensions are likely to become increasingly common as healthcare institutions begin to evaluate exploratory generative AI projects using governance frameworks originally designed for conventional biomedical research.

Acknowledgments

This work is supported by the project “Responsible AI in Healthcare: Integrating Language Models, Clinical Knowledge, and Ethical Evaluations” (FMV_1_2025_1_186830), funded by the Maria Viñas Fund of the National Agency for Research and Innovation (ANII), Uruguay.

The authors would like to thank the other researchers in the project for their valuable contributions during the project design: Lucía Tort (Hospital de Clínicas, UDELAR); Aiala Rosá, Guillermo Moncecchi, and Diego Garat (InCo, Facultad de Ingeniería, UDELAR); María Goñi (CSIC, UDELAR); and María Noelia Beltramelli (CICADA, UDELAR). We also thank Daniel Taborda (FCM - UNC) and Guido Ivetta (FAMAF - UNC) for their essential role in the development of the simulated virtual patients project.

We are also grateful to the High Performance Computing Center (CCAD) of the Universidad Nacional de Córdoba (<https://ccad.unc.edu.ar/>), part of SNCAD-MinCyT, Argentina. Although the present study did not directly use its computational infrastructure, CCAD hosts the open-weight LLMs used in our simulated virtual patients project.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to assist with grammar and spelling checks. After using this tool, the bilingual authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] D. S. Char, N. H. Shah, D. Magnus, Implementing machine learning in health care—addressing ethical challenges, *The New England journal of medicine* 378 (2018) 981. doi:10.1056/NEJMp1714229.
- [2] J. Morley, C. C. Machado, C. Burr, J. Cows, I. Joshi, M. Taddeo, L. Floridi, The ethics of AI in health care: a mapping review, *Social science & medicine* 260 (2020) 113172. doi:10.1016/j.socscimed.2020.113172.
- [3] R. Hatem, B. Simmons, J. E. Thornton, A call to address AI “hallucinations” and how healthcare professionals can mitigate their risks, *Cureus* 15 (2023). doi:10.7759/cureus.44720.
- [4] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, E. Cambria, A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics, *Information Fusion* 118 (2025) 102963. doi:10.1016/j.inffus.2025.102963.
- [5] M. Fareed, M. Fatima, J. Uddin, A. Ahmed, M. A. Sattar, A systematic review of ethical considerations of large language models in healthcare and medicine, *Frontiers in Digital Health* 7 (2025) 1653631. doi:10.3389/fdgth.2025.1653631.
- [6] Y. Ji, W. Ma, S. Sivarajkumar, H. Zhang, E. M. Sadhu, Z. Li, X. Wu, S. Visweswaran, Y. Wang, Mitigating the risk of health inequity exacerbated by large language models, *npj Digital Medicine* 8 (2025) 246. doi:10.1038/s41746-025-01576-4.
- [7] B. Mittelstadt, Principles alone cannot guarantee ethical AI, *Nature machine intelligence* 1 (2019) 501–507. doi:10.1038/s42256-019-0114-4.
- [8] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, W. R. Hersh, A systematic literature review of automated clinical coding and classification systems, *Journal of the American Medical Informatics Association* 17 (2010) 646–651. doi:10.1136/jamia.2009.001024.
- [9] H. Dong, M. Falis, W. Whiteley, B. Alex, J. Matterson, S. Ji, J. Chen, H. Wu, Automated clinical coding: what, why, and where we are?, *npj Digital Medicine* 5 (2022) 159. doi:10.1038/s41746-022-00705-7.
- [10] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at

- CodiEsp track of CLEF eHealth 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696 of *CEUR Workshop Proceedings*, 2020.
- [11] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: Models and evaluation metrics, *Journal of the American Medical Informatics Association* 21 (2014) 231–237. doi:10.1136/amiajnl-2013-002159.
 - [12] M. Falis, H. Dong, A. Birch, B. Alex, CoPHE: A count-preserving hierarchical evaluation metric in large-scale multi-label text classification, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 907–912. doi:10.18653/v1/2021.emnlp-main.69.
 - [13] M.-I. Chen, Y.-L. Hsu, Y.-H. Chen, Revenue-sensitive evaluation of AI-assisted ICD-10-CM coding and human-AI collaboration under dual DRG payment systems, *Scientific Reports* 16 (2026) 26729. doi:10.1038/s41598-026-57682-0.
 - [14] T. Searle, Z. Ibrahim, R. Dobson, Experimental evaluation and development of a silver-standard for the MIMIC-III clinical coding dataset, in: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 2020, pp. 76–85. doi:10.18653/v1/2020.bionlp-1.8.
 - [15] S. Khadka, X. Jiang, V. Palade, Llm-assisted clinical coding audit through an interpretable coding pipeline, *International Journal of Medical Informatics* 220 (2026) 106605. doi:https://doi.org/10.1016/j.ijmedinf.2026.106605.
 - [16] S. Li, C. Zheng, J. Wu, Q. Xu, X. Xu, H. Wang, Y. Sun, Z. Bai, Y. Xu, L. Zhu, W. Hu, F. Huang, Verification is all you need: Prompting large language models for zero-shot clinical coding, *IEEE Journal of Biomedical and Health Informatics* 29 (2025) 8536–8549. doi:10.1109/JBHI.2025.3593028.
 - [17] J. H. Aranda, S. M. Monks, Roles and Responsibilities of the Standardized Patient Director in Medical Simulation, in: *StatPearls*, StatPearls Publishing, Treasure Island (FL), 2026.
 - [18] D. Bauer, F.-M. Lahner, S. Huwendiek, F. M. Schmitz, S. Guttormsen, An overview of and approach to selecting appropriate patient representations in teaching and summative assessment in medical education, *Swiss Medical Weekly* 150 (2020) w20382. doi:10.4414/smw.2020.20382.
 - [19] J. Kaplonyi, K.-A. Bowles, D. Nestel, D. Kiegaldie, S. Maloney, T. Haines, C. Williams, Understanding the impact of simulated patients on health care learners' communication skills: a systematic review, *Medical Education* 51 (2017) 1209–1219. doi:10.1111/medu.13387.
 - [20] F. Holderried, C. Stegemann-Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. Festl-Wietek, M. Mahling, A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study, *JMIR Medical Education* 10 (2024) e53961. doi:10.2196/53961.
 - [21] J. J. Webb, Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News, *Cureus* 15 (2023) e38755. doi:10.7759/cureus.38755.
 - [22] C. Wang, S. Li, N. Lin, X. Zhang, Y. Han, X. Wang, D. Liu, X. Tan, D. Pu, K. Li, G. Qian, R. Yin, Application of Large Language Models in Medical Training Evaluation—Using ChatGPT as a Standardized Patient: Multimetric Assessment, *J Med Internet Res* 27 (2025) e59435. doi:10.2196/59435.
 - [23] G. Ivetta, M. J. Gomez, S. Martinelli, P. Palombini, M. E. Echeveste, N. C. Mazzeo, B. Busaniche, L. Benotti, HESEIA: A community-based dataset for evaluating social biases in large language models, co-designed in real school settings in Latin America, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Suzhou, China, 2025, pp. 25095–25117. doi:10.18653/v1/2025.emnlp-main.1275.
 - [24] P. Brown, S. C. Levinson, *Politeness: Some universals in language usage*, Cambridge university press, 1987.
 - [25] S. Martinelli, D. E. Taborda, G. Ivetta, L. Benotti, Assessment of llm-based simulated virtual patients in clinical simulations of conversations, in: *Workshop on Ethics in Clinical AI Applications*

(ETHICAIA 2026), Bremen, Germany, 2026. Forthcoming.

- [26] D. Leslie, Understanding artificial intelligence ethics and safety, The Alan Turing Institute (2019). doi:10.5281/zenodo.3240529.
- [27] D. Martinho, P. Sobreiro, A. Domingues, F. Martinho, N. Nogueira, Ethical responsibility in medical ai: A semi-systematic thematic review and multilevel governance model, Healthcare 14 (2026) 287. doi:10.3390/healthcare14030287.