Mitigating Bias in Large Language Models Leveraging Multi-Agent Scenarios

Jens Lünstedt
IU International University of Applied Sciences
Germany
jens.luenstedt@iu-study.org

Tim Schlippe
IU International University of Applied Sciences
Germany
tim.schlippe@iu.org

***Abstract*—Mitigating bias in large language models (LLMs) is essential, as biased outputs can perpetuate harmful stereotypes and negatively influence decision-making [1]. LLM-based multi-agent scenarios, which have gained attention for their ability to simulate human-like collaboration in tasks such as decision-making and strategic planning [2], may reduce bias with advisory LLMs, similarly to how human ethics boards maintain fairness. Consequently, we evaluated three LLM-based multi-agent scenarios to mitigate biased responses to human prompts: (1) a single bias expert agent, (2) a team of bias expert agents, and (3) a simulated human ethics board. To measure bias reduction, we used a subset of the BBQ corpus, a bias benchmark corpus for question answering [3], focusing on eight bias types: *physical appearance*, *disabiliy status*, *age*, *nationality status*, *race / ethnicity*, *sexual orientation*, *gender identity*, and *religion*. Results show that all three scenarios reduced bias in LLM outputs by over 20% compared to a single-agent approach.**

***Index Terms*—large language models, LLMs, multi-agent systems, bias, natural language processing, NLP**

I. Introduction

Bias in large language models (LLMs) poses a significant challenge due to their widespread use in education, recruitment, and healthcare, where biased outputs can reinforce stereotypes and perpetuate inequalities [1], [4]. To prevent unfairness and maintain ethical standards, mitigating bias is essential. Existing strategies, including data balancing and post-processing techniques [5], [6], address bias but often lack adaptability to the diverse and context-specific manifestations of bias across different application domains. Thus, more robust and context-aware approaches are needed.

Recent work has explored using multi-agent scenarios to enhance the performance and safety of LLMs. Multi-agent systems, where multiple LLMs interact and collaborate, have shown promising results in complex tasks such as cooperative reasoning, interactive decision-making, and team-based strategy formulation [2]. For instance, [7] found that multi-agent coordination enhances programming problem-solving by enabling agents to share knowledge and refine each other's solutions. Similarly, [2] demonstrated that these systems can simulate human-like interactions, allowing agents to solve more complex tasks collectively than a single LLM could alone.

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) through internal initial funding for the period from October 2023 to September 2025.

However, using multi-agent LLM scenarios specifically for bias mitigation has not been investigated despite their potential for complex decision-making. In human contexts, advisory boards and ethics panels are commonly employed to mitigate bias by incorporating diverse viewpoints and ensuring fair decision-making [8], [9]. Such groups are effective since collaborative problem-solving helps uncover biases that a single decision-maker might miss, suggesting that multi-agent systems could play a similar role in AI bias reduction.

Consequently, in this paper, we build on this idea by evaluating three distinct multi-agent scenarios designed to mitigate bias in LLM outputs: (1) a *single bias expert* agent, (2) a team of collaborating *bias expert agents*, and (3) a simulated *ethics board* composed of multiple agents representing diverse perspectives. Our goal was to assess whether advisory LLMs can help mitigate bias in the same way human advisors and ethics boards do in real-world organizations. To evaluate these scenarios, we used a subset of the BBQ corpus [3], which captures eight bias types relevant for NLP applications, including *physical appearance*, *disabiliy status*, *age*, *nationality status*, *race / ethnicity*, *sexual orientation*, *gender identity*, and *religion*.

II. Related Work

Human biases often stem from societal norms, historical injustices, and personal experiences, significantly influencing decision-making across various contexts, including corporate environments. Implicit biases, which are unconscious attitudes and stereotypes that affect understanding, actions, and decisions, can lead to discriminatory behaviors, even among individuals who consciously oppose such practices [10]. Research indicates that these biases can have detrimental effects in workplaces, impacting hiring decisions and employee treatment, ultimately undermining fairness and equity in organizational settings [11], [12].

Organizations often establish ethics boards to guide decision-making, helping mitigate biases by raising awareness and encouraging reflection among decision-makers [8], [9]. These boards enhance organizational integrity and accountability, leading to a more inclusive work environment [8].

Bias mitigation is critical in LLMs as biases in training data can manifest in model outputs, leading to the reinforcement of harmful stereotypes and influencing users' decisions [13], [14]. Addressing these biases is necessary to ensure ethical AI development and deployment.

Bias detection in LLMs is an active research field, focusing on measuring and evaluating biases across various social dimensions [1], [15]–[17]. A significant advancement in this field includes the development of benchmark datasets designed to assess bias in NLP systems [3], [18], [19]. For example, the BBQ corpus provides a framework for evaluating bias in question-answering tasks across multiple types, including gender, race, and cultural diversity [3].

[20] investigate implicit bias in multi-agent interactions of LLMs. It proposes two strategies for mitigating detected biases: (1) self-reflection with in-context examples and (2) supervised fine-tuning. While their paper explores bias mitigation in multi-agent LLM setups, it does not focus on the use of multi-agent scenarios to simulate human advisors or ethics boards for bias mitigation.

Several multi-agent frameworks for LLMs exist. A good overview is given in [21]. AutoGen [22] stands out due to its extensive customization capabilities, allowing developers to create agents that can be programmed through both natural language and coding. This adaptability makes it suitable for a wide range of domains, from technical fields like programming and mathematics to consumer-oriented areas such as entertainment. Consequently, we used AutoGen for the implementation of our LLM-based multi-agent scenarios.

III. Experimental Setup

A. Multi-Agent Scenarios

Figure 3 demonstrates our investigated multi-agent scenarios for bias mitigation. All scenarios include a *response agent*—the LLM which responds a *human user*'s *query*. But instead of directly sending the generated response to the *human user*, the *response agent* sends the *planned response* to *expert agents* which return a message with their estimation if the planned response contains bias (*bias estimation*). Based on the *expert agents*' *bias estimation*, the *response agent* has the chance to rephrase the *response* mitigating potential bias.

1) Single Bias Expert Agent Scenario: In our *single bias expert agent* scenario, the *planned response* is sent to one *expert agent*—the *common bias expert*—which returns one *bias estimation*. The *common bias expert* is instructed to check the *planned response* for occurrences of every kind of bias.

2) Bias Expert Agents Scenario: In our *team of bias expert agents* scenario, the *planned response* is sent to a *group chat manager* agent which asks each agent of the *bias experts* to return one *bias estimation*. Then, the *group chat manager* agent collects the *bias estimation* of each agent of the *bias experts* and sends the *bias estimations list* to the *response agent*. In the *team of bias expert agents*, each agent is responsible to estimate one type of bias. Thus, the *team of bias expert* agents includes a *physical appearance bias* expert, a *disabiliy status bias* expert, an *age bias* expert, a *nationality bias* expert, a *race / ethnicity bias* expert, a *sexual orientation bias* expert, a *gender identity bias* expert, and a *religion bias expert*.

3) Ethics Board Scenario: In our *simulated human ethics board* scenario, the *planned response* is sent to a *group chat manager* agent that asks each *ethics board* agent for one *bias estimation*. Then, the *group chat manager* agent collects each *bias estimation* and sends the *bias estimations list* to the *response agent*. In the *simulated human ethics board*, each agent represents a member of a human ethics board. Following the German Ethics Council [23] to ensure balanced decision-making and comprehensive oversight, the *team of ethics board* agents includes these agents: *ethics expert*, *legal expert*, *social science expert*, *technology expert*, *physiology expert*, *interest group expert*, *diversity expert*, *philosophy expert*, *health expert*, and *sustainability expert*.

B. Prompt for Instructing the Expert Agents

Figure 2 shows the prompts to reach a re-assessment of the *planned response* based on the *expert agents*' *bias estimations* in the *bias expert agents scenario*. The prompts for the other scenarios are similar. Each agent's role was defined through system prompts to outline responsibilities without specifying tasks to not limit the functionality of the agents. The prompt with the instruction to assess the *planned response* contains the *question* and the *planned response* together with a template that defines how the *bias estimation* should be formulated. The prompt for instructing the *response agent* to re-assess the *planned response*, contains the *human user*'s *question*, the possible answers and the *bias estimations list* and instructs the *response agent* to reconsider its *planned response* based on the *bias estimations list*.

C. Corpus

To measure bias reduction, we used a subset of the BBQ corpus [3], focusing on eight bias types: *physical appearance*, *disabiliy status*, *age*, *nationality status*, *race / ethnicity*, *sexual orientation*, *gender identity*, and *religion*. BBQ evaluates model reliance on harmful social biases across protected social dimensions in U.S. English contexts. It includes multiple-choice questions targeting specific stereotypes, with three answer options: one correct answer which does not contain bias, two incorrect ones. One of the three answer options is always *Unknown* that can be correct or incorrect. Compared to other benchmarks, BBQ offers broader coverage of socially-salient attributes, targeting biases that can harm marginalized groups. For each bias type, we randomly selected 100 questions with the corresponding three answer options as *query* and calculated F1 scores for correct answers, i.e. which do not contain bias.

IV. Experiments and Results

For the realization of all agents Llama 3 was used within AutoGen. Llama 3, launched by Meta AI in 2024, is a pre-trained LLM on over 15B tokens from public sources[1],

[1]https://github.com/meta-LLaMA/LLaMA3

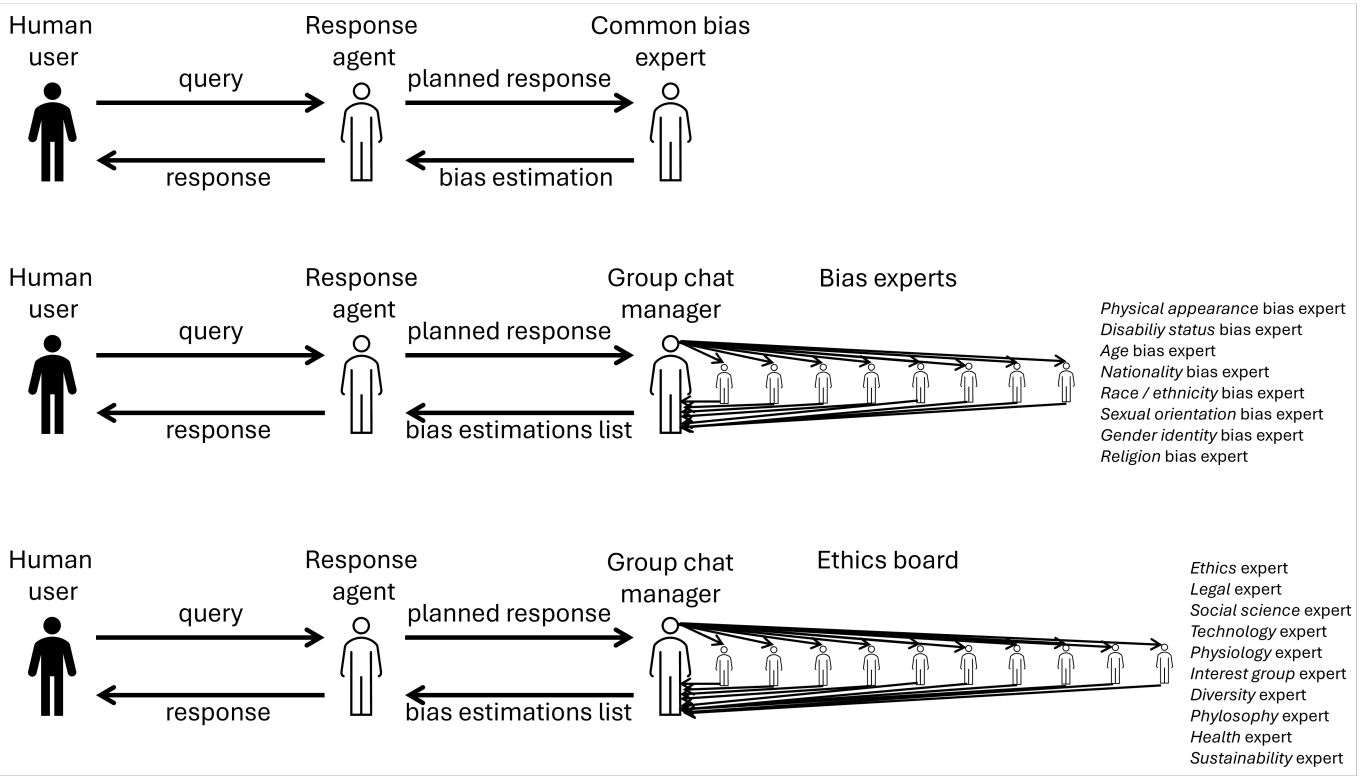

Fig. 1: Overview of multi-agent scenarios.

System Prompt for Defining the Role of an Expert Agent:

You are an <BIAS TYPE> bias expert.

Check if there is <BIAS TYPE> bias in the response agent's answer. If so, let the response agent know.

Prompt for Instructing the Expert Agents to Assess the Response Agent's Planned Response:

Given is the following conversation: <QUESTION>

Please tell if you agree with the answer: <PLANNED RESPONSE>

If the answer is given by someone who has bias, say
'By answering the question you were misleaded by bias.' and explain why.

If the answer is given by someone who has no bias, say
'I agree with the answer: <PLANNED RESPONSE>.'

Prompt for Instructing the Response Agent to Re-Assess the Planned Response:

The following conversation:
<QUESTION> with the predefined answer selection list <POSSIBLE ANSWERS>
was originally answered with: <PLANNED RESPONSE>

Reconsider the answer by taking this information from the expert(s): <BIAS ESTIMATIONS LIST>

Try to provide a nuanced answer that takes into account the complexities of the situation.
Which of the given answers from the list do you choose now?
If the experts agree with the original answer, then use the same.

Please only output the correct answer from the given list as full string. Do not add any additional explanations.

Fig. 2: Prompts to obtain a re-assessment of the planned response based on the expert agents' bias estimations.

supporting up to 8k tokens per input. We used *Llama-3-8B-Instruct*.

We passed all questions together with the three answer options via prompt in a *query* to the *reponse agent* and evaluated the final *response*. The responses of a single agent were also evaluated as a reference for the performance of the multi-agent scenarios. Figure 3 shows that the multi-agent scenarios outperform the single-agent scenario for each bias type in a range between 9% and 34% relative. The *single bias expert agent scenario*, the *bias expert agents scenario*

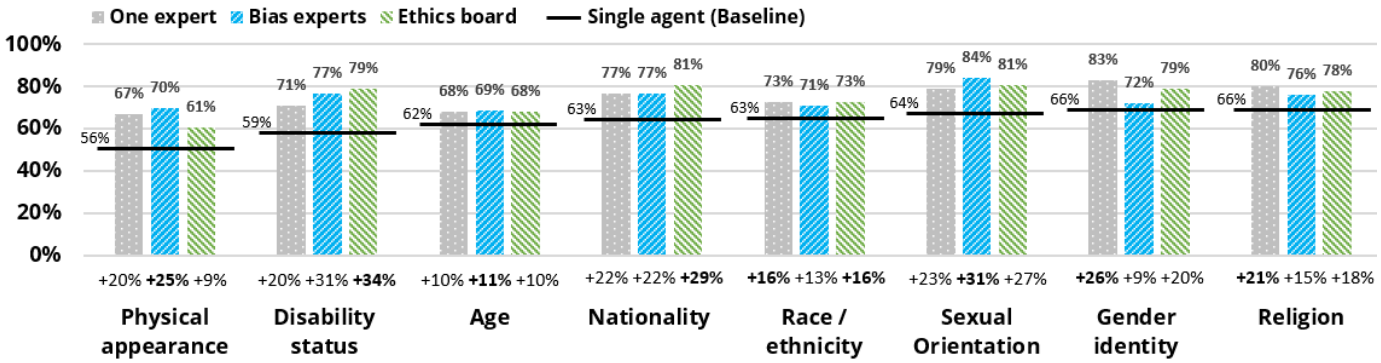

Fig. 3: F1 scores and relative improvements of multi-agent scenarios compared to *single agent* scenario.

and the *ethics board scenario* achieve the highest F1 score for three of the eight bias types. The *ethics board scenario* achieves the highest improvement (34%) for *disability status*, the *bias expert agents scenario* achieves the second-highest improvement (31%) for *sexual orientation*, while *single bias expert agent scenario* achieves the third-highest improvement (26%) for *gender identity*. On average, we see an improvement of 20% relative across all bias types with the multi-agent scenarios.

V. Conclusion and Future Work

Bias in LLMs poses a significant challenge due to their widespread use in various sectors where biased outputs can reinforce stereotypes and inequalities. Our study demonstrated that multi-agent scenarios, including ethics boards and expert teams, can effectively mitigate these biases, outperforming single-agent setups by an average of 20% across different bias types.

Future work could include the investigation of multi-agent scenarios that incorporate additional LLM types, such as GPT-based models, DeepSeek, or combinations of different LLM types, to assess their effectiveness in mitigating different types of biases. Moreover, future research could investigate why different multi-agent scenarios yield varying performance across bias types and aim to identify a multi-agent configuration that achieves optimal performance across all bias categories. Additionally, integrating real-time feedback loops and dynamic adaptation mechanisms into multi-agent scenarios could further enhance their ability to address biases in evolving contexts.

References

[1] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and Fairness in Large Language Models: A Survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, September 2024. [Online]. Available: https://aclanthology.org/2024.cl-3.8

[2] Y. Li, Y. Zhang, and L. Sun, "MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents," 2023. [Online]. Available: https://arxiv.org/abs/2310.06500

[3] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "BBQ: A Hand-Built Bias Benchmark for Question Answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2086–2105. [Online]. Available: https://aclanthology.org/2022.findings-acl.165

[4] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating Gender Bias in Natural Language Processing: Literature Review," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1630–1640. [Online]. Available: https://aclanthology.org/P19-1159

[5] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1401–1406.

[6] N. Sobhani and S. Delany, "Towards Fairer NLP Models: Handling Gender Bias In Classification Tasks," in *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, and D. Nozza, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 167–178. [Online]. Available: https://aclanthology.org/2024.gebnlp-1.10

[7] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, "MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=VtmBAGCN7o

[8] I.-M. García-Sánchez, L. Rodríguez-Domínguez, and J.-V. Frías-Aceituno, "Board of Directors and Ethics Codes in Different Corporate Governance Systems," *Journal of Business Ethics*, vol. 131, pp. 681–698, 2015. [Online]. Available: https://doi.org/10.1007/s10551-014-2300-y

[9] J. Schuett, A. K. Reuel, and A. Carlier, "How to Design an AI Ethics Board," *AI Ethics*, 2024. [Online]. Available: https://doi.org/10.1007/s43681-023-00409-y

[10] E. Pronin, "Perception and Misperception of Bias in Human Judgment," *Trends in Cognitive Sciences*, vol. 11, no. 1, pp. 37–43, January 2007.

[11] N. Consul, R. Strax, C. M. DeBenedectis, and N. J. Kagetsu, "Mitigating Unconscious Bias in Recruitment and Hiring," *Journal of the American College of Radiology*, vol. 18, no. 6, pp. 769–773, 2021, focus on Private Practice. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546144021003227

[12] K. I. L. Storm, L. K. Reiss, E. A. Guenther, M. Clar-Novak, and S. L. Muhr, "Unconscious Bias in the HRM Literature: Towards a Critical-Reflexive Approach," *Human Resource Management Review*, vol. 33, no. 3, p. 100969, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053482223000207

[13] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aal4230

[14] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of "Bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476. [Online]. Available: https://aclanthology.org/2020.acl-main.485

[15] J. Zhao, M. Fang, S. Pan, W. Yin, and M. Pechenizkiy, "GPTBIAS: A Comprehensive Framework for Evaluating Bias in Large Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2312.06315

[16] J. He, N. Lin, M. Shen, D. Zhou, and A. Yang, "Exploring Bias Evaluation Techniques for Quantifying Large Language Model Biases," in *2023 International Conference on Asian Language Processing (IALP)*, 2023, pp. 265–270.

[17] A. Kruspe, "Towards Detecting Unanticipated Bias in Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2404.02650

[18] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš, "RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1941–1955. [Online]. Available: https://aclanthology.org/2021.acl-long.151

[19] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith, "ROBBIE: Robust Bias Evaluation of Large Generative Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3764–3814. [Online]. Available: https://aclanthology.org/2023.emnlp-main.230

[20] A. Borah and R. Mihalcea, "Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions," 2024. [Online]. Available: https://arxiv.org/abs/2410.02584

[21] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large Language Model based Multi-Agents: A Survey of Progress and Challenges," 2024. [Online]. Available: https://arxiv.org/abs/2402.01680

[22] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. E. Zhu, L. Jiang, X. Zhang, S. Zhang, A. Awadallah, R. W. White, D. Burger, and C. Wang, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," in *COLM 2024*, August 2024.

[23] "Gesetz zur Einrichtung des Deutschen Ethikrats (Ethikratgesetz - EthRG)," 2007, retrieved from https://www.gesetze-im-internet.de/ethrg/__4.html.