Generative AI: Developments, Applications, and Responsible Practices

Alberto Moccardi[1,*], Egidia Cirillo[1], Cristina Davino[2], Mattia Fonisto[1], Francesco Gargiulo[4], Rajib Chandra Ghosh[1], Ojasvi Gupta[3], Rajesh Jaiswal[3], Roberto La Rovere[1], Lidia Marassi[1], Zahida Mashaallah[1], Narendra Patwardhan[1], Gian Marco Orlando[1], Domenico Benfenati[1], Giovanni Maria De Filippis[1], Antonio Elia Pascarella[1], Diego Russo[5], Cristiano Russo[1], Cristian Tommasino[1], Stefano Marrone[1,*], Flora Amato[1], Antonio Maria Rinaldi[1], Vincenzo Moscato[1] and Carlo Sansone[1]

[1]*Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy,*

[2]*Department of Economics and Statistics (DiSES), University of Naples Federico II, Via Cintia 21, 80126 Naples, Italy,*

[3]*Technological University Dublin, Dublin, Ireland,*

[4]*National Research Council (CNR), Italy,*

[5]*Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy*

Abstract

Generative AI significantly influences various sectors, including education, communication, legal decision-making, and academia. In education, AI platforms, such as Sofia, deliver personalized and multilingual support, enhancing inclusivity despite ethical and accessibility concerns. Furthermore, Retrieval-Augmented Generation (RAG) approaches in scientific domains effectively reduce the limitations of large language models (LLMs) by dynamically integrating trusted scientific information, thereby improving the accuracy and reliability of responses. However, critical issues continue regarding inherent social biases within LLMs, notably related to gender, ethnicity, disability, and disinformation. In response to these limitations, Human-in-the-Loop (HITL) procedures are presented here, combining high interpretability and human oversight, thus enhancing AI systems' accuracy and ethical compliance in the legal domain. Accordingly, generative agents powered by LLMs that emulate human behaviors are presented as powerful tools for unbiased participation in information verification tasks. Overall, this article presents responsible and ethically aligned practices essential for leveraging generative AI's transformative potential across multiple domains and sectors.

Keywords

Artificial Intelligence, Generative AI, Large Language Models

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

* Corresponding author.

alberto.moccardi@unina.it (A. Moccardi); egidia.cirillo@unina.it (E. Cirillo); cristina.davino@unina.it (C. Davino); mattia.fonisto@unina.it (M. Fonisto); francesco.gargiulo@icar.cnr.it (F. Gargiulo); rajib.chandraghosh@unina.it (R. C. Ghosh); ojasvi.gupta@tudublin.ie (O. Gupta); rajesh.jaiswal@tudublin.ie (R. Jaiswal); r.larovere88@gmail.com (R. L. Rovere); lidia.marassi@unina.it (L. Marassi); zahida.mashaallah@unina.it (Z. Mashaallah); narendraprakash.patwardhan@unina.it (N. Patwardhan); gianmarco.orlando@unina.it (G. M. Orlando); domenico.benfenati@unina.it (D. Benfenati); giovannimaria.defilippis@unina.it (G. M. D. Filippis); antonioelia.pascarella@unina.it (A. E. Pascarella); diego.russo@unibg.it (D. Russo); cristiano.russo@unina.it (C. Russo); cristian.tommasino@unina.it (C. Tommasino); stefano.marrone@unina.it (S. Marrone); flora.amato@unina.it (F. Amato); antoniomaria.rinaldi@unina.it (A. M. Rinaldi); vincenzo.moscato@unina.it (V. Moscato); carlo.sansone@unina.it (C. Sansone)

0009-0001-6136-93688 (A. Moccardi); 0009-0005-3227-6073 (E. Cirillo); 0000-0003-1154-4209 (C. Davino); 0000-0002-2422-0425 (M. Fonisto); 0000-0003-0400-3332 (F. Gargiulo); 0009-0001-1137-3465 (R. C. Ghosh); 0009-0003-5159-0004 (O. Gupta); 0000-0002-4530-7079 (R. Jaiswal); 0009-0006-8134-5466 (L. Marassi); 0000-0002-4807-5664 (N. Patwardhan); 0009-0008-5825-8043 (D. Benfenati); 0009-0002-8395-0724 (G. M. D. Filippis); 0000-0002-1079-7741 (A. E. Pascarella); 0000-0002-8732-1733 (C. Russo); 0000-0001-9763-8745 (C. Tommasino); 0000-0001-6852-0377 (S. Marrone); 0000-0002-5128-5558 (F. Amato); 0000-0001-7003-4781 (A. M. Rinaldi); 0000-0002-4807-5664 (V. Moscato); 0000-0002-8176-6950 (C. Sansone)

1. Generative AI and Education

In recent years, continuous technological progress has led to the expansion of digital solutions across multiple domains, with education emerging as a key area of application. Artificial Intelligence (AI) has become a valuable resource for improving educational support by simplifying access to essential information, enabling personalized learning, and fostering greater inclusion. An illustrative example is **Sofia**[1], an AI-based platform designed to enhance the university experience. Sofia demonstrates how AI can streamline academic processes and improve student services, reflecting the increasing influence of intelligent systems on the educational landscape. Tools like Sofia, powered by AI, are proving to be highly effective in assisting the academic community. Technologies such as natural language models and conversational agents are revolutionizing the way students interact with institutions, offering tailored, on-demand assistance that requires no technical expertise and operates independently of time constraints. A notable strength of AI-based chatbots like Sofia is their ability to deliver timely, accurate responses to specific inquiries, thereby enhancing access to information for diverse user groups. In academic settings, such capabilities help eliminate informational barriers, facilitating student access to critical data ranging from administrative matters to educational resources. These tools ensure students receive continuous, targeted support throughout their academic journey.

Another major benefit of AI integration in education is its capacity to embrace linguistic and cultural diversity. In Sofia's case, the system's multilingual functionality significantly broadens its reach, making it a valuable resource for international students and supporting a more inclusive and customized learning environment. While Sofia stands as a concrete example of AI supporting student guidance, many other applications exist for adapting education to individual student needs. AI technologies can be embedded across the entire academic lifecycle, offering support not only during orientation but throughout the student's educational progression.

Overall, AI is poised to become a central force not only in reshaping education but also in influencing broader societal change. As such, understanding and managing the human-machine relationship may itself become a vital educational pursuit in the years ahead [1].

2. Social Biases in Large Language Models

In this section, we describe our empirical investigation into how social biases manifest in large language models (LLMs), with a specific focus on biases related to gender, ethnicity, and disability. As LLMs such as Mistral, LLAMA, and Gemma become increasingly integrated into systems used in healthcare, employment, public services, and digital interfaces, it is essential to examine their fairness properties critically. These models, trained on vast and often uncurated corpora, frequently inherit and reproduce social prejudices, either explicitly or in more subtle, systemic ways [2].

We characterize bias in LLMs as either intrinsic or extrinsic. Intrinsic bias refers to latent tendencies embedded in the model's learned parameters, observable across general prompts and use cases. Extrinsic bias, by contrast, emerges in downstream tasks and applications, often modulated by context, prompt formulation, or task framing [3]. The core of our study is structured around three research questions: (1) are LLMs more susceptible to bias through direct versus indirect adversarial prompts? (2) does explicit context in prompts meaningfully reduce biased responses? and (3) do model optimization strategies, such as quantization, affect the emergence of bias?

To explore these questions, we implemented a structured prompt-based testing pipeline. Prompts were drawn from both direct stereotype datasets (GEST) [4] and contextual question-answering datasets (BBQ) [5], focusing on English-language entries relevant to the three targeted bias categories. Each prompt was evaluated on three open-source LLMs: Mistral 7B, LLAMA 2 7B, and Gemma 2 9B. Where available, both base pre-trained and fine-tuned model variants were used. All models were accessed through the HuggingFace Transformers interface, allowing for reproducible experimentation.

[1] https://sofia.dises.ai/

Our results show that LLMs demonstrate high sensitivity to directly biased prompts. Approximately 45% of responses to direct prompts displayed stereotypical assumptions, especially in the gender domain. Fine-tuned models exhibited varied behaviors. For instance, the fine-tuned Mistral variant produced fewer overtly biased outputs but increased the proportion of "confused" or evasive responses, which we define as noncommittal or self-censoring. In contrast, the fine-tuned LLAMA model tended to respond more decisively but occasionally reinforced biased reasoning. Prompts relating to disability, gender, and ethnicity elicited the most biased responses across all model types,particularly in no-context conditions, suggesting that this category remains underrepresented or misrepresented in training corpora [6].

Overall, our findings indicate that while instruction tuning and context-awareness mechanisms improve surface-level safety, they do not resolve deeper representational biases, particularly in the absence of specific contextual disambiguation.

3. Human-in-the-Loop Generative AI in Legal Domain

The Human-in-the-Loop (HITL) process is an interactive approach to generative artificial intelligence that enhances the precision and explainability of complex generative architectures by incorporating human insights into model learning and addressing model oversights during the inference process, a key feature of legal decision support systems (DSS).

The application of HITL strategies in AI-driven legal assistants aligns with the indications of the European AI Act, which aims to ensure the compliance of AI-driven systems with regulations and ethical standards. In this context, the study by Amatrian [7], elucidates different prompting-based HITL methodologies, focusing mainly on Reasoning without Observation (ReWOO), Reasoning and Action (ReAct), and Dialog-Enabled Resolving Agents (DERA) paradigms.

DERA is a general chat framework that leverages dialog-capable agents to work through a task iteratively. This method consists of two agents: (i) the researcher and (ii) the decider. In particular, the researcher parallels human decision-making by enabling complex queries to be handled thoroughly.

ReAct is a novel prompt-based paradigm that synergizes reasoning and acting in language models for general task solving. The general workflow can be grouped into two interconnected parts: reasoning and action. The agent searches for useful assets within the user query and then generates thoughts addressing the right tools to find these assets within the external legal knowledge base. Then, interrogating the right tools that refer to different portions of the knowledge base, relevant content for the thought is retrieved for answer formulation, as in a classical RAG-based system.

ReWOO, which is built on ReAct, compartmentalizes the workflow into three separate modules: (i) a planner, (ii) the workers, and (iii) a solver. The planner breaks down a task and formulates a blueprint of interdependent plans; each allocated to a worker. The workers retrieve external knowledge from tools to provide evidence. The solver synthesizes the plans to generate the ultimate answer.

This high-level comparison highlights the superior capabilities of ReAct-based HITL systems, as they offer outstanding human oversight and control over internal generative mechanisms, ensuring adherence to technological standards, user needs, and professional expectations within the legal domain.

4. Generative Agents for Social Simulation and Crowdsourced Fact-Checking

Over the past decades, there has been a concerted effort among researchers and practitioners to develop computational agents capable of realistically emulating human behavior [8]. Within this context, Agent-Based Modelling (ABM) has emerged as a pivotal methodology for simulating complex systems by defining rules that govern individual agents' behavior and interactions [9].

Recent advancements in Large Language Models (LLMs) have enabled the emergence of *generative agents*, autonomous computational entities capable of reasoning, adapting, and interacting in ways resembling human behavior. Generative Agent-Based Modeling (GABM) leverages these capabilities

and integrates LLM with traditional ABM to produce agents whose behavior arises dynamically from interactions within the environment, overcoming the limitations of rule-based approaches [10, 11, 12].

We have contributed to this research direction by developing and validating a GABM framework for simulating social media environments using real-world network data[2]. Our initial focus was on assessing the fidelity of generative agents in replicating user-level properties, such as linguistic style, interests, and political orientation, derived from large-scale Twitter datasets. To this end, we developed a modular simulation pipeline comprising: (i) an Agent Characterization Module for encoding user-specific personality traits and interests; (ii) a Reasoning Module powered by LLMs to determine agent actions (e.g., posting, re-sharing, remaining silent); and (iii) an Interaction Module that governs content exposure through different recommendation strategies. Experimental results showed that LLM-agents effectively preserve the ideological alignment and communication styles of the real users they emulate, while also reproducing emergent phenomena such as homophily and polarization [13].

Beyond simulation contexts, we explored the use of generative agents in a distinct, yet socially impactful application: crowdsourced fact-checking [14]. In this setting, generative agents were tasked with impersonating realistic user profiles—defined by demographic and ideological attributes—and participating in structured fact-verification workflows. Our results indicate that agent-based crowds not only outperformed human annotators in classifying claim veracity, but also, unlike human contributors, maintained stable performance across demographic and ideological variations, suggesting their potential to mitigate biases commonly observed in human-based crowdsourcing.

5. RAG for scientific documents

To overcome the inherent limitations of standard large language models (LLMs)—such as hallucinations, limited factual grounding, and context length constraints—we developed a Retrieval-Augmented Generation (RAG) framework specifically designed for tasks involving scientific literature and domain-specific knowledge integration [15, 16, 17]. Rather than relying solely on pre-trained model weights, our RAG architecture dynamically retrieves relevant information from a curated collection of scientific documents, thereby injecting real-time factual context into the generation process. This approach is particularly effective in specialized domains where precision and traceability are critical, such as biomedical and nutritional sciences [18]. The backbone of our system is a vector database (VDB) architecture that indexes scientific texts—ranging from peer-reviewed articles to curated metadata—using dense vector embeddings. These embeddings, generated through domain-optimized encoders such as General Text Embeddings (GTE) [19], allow the system to compute semantic similarity between user queries and document passages via cosine similarity measures. Upon receiving a query, the RAG pipeline retrieves the top-ranked passages most semantically aligned with the input, which are then used to condition the generation of the final response. This real-time retrieval mechanism anchors the language model's output in verifiable sources, mitigating the risk of hallucinations and improving the factual reliability of responses [20, 21]. Our research focuses on applying this RAG framework to a corpus of scientific literature, particularly in domains like nutrigenetics, where terminology is complex and answers require precise contextual grounding. By using peer-reviewed publications and structured metadata as the retrieval base, we ensure that the model operates with a high degree of factual alignment and scientific validity. Our RAG-enhanced pipeline significantly outperforms baseline models such as GPT-3.5 and Mistral-7B across relevance, accuracy, and specificity metrics. These results underscore the effectiveness of integrating scientific documents into the RAG paradigm, enabling LLMs to function as reliable tools for scientific reasoning.

6. Programmatic Agency

The development of Hominis centers on advancing generative model capabilities through robust pretraining, sophisticated instruction tuning, and a novel agentic framework for reasoning and behavior

[2]The framework is available at: https://github.com/PRAISELab-PicusLab/LLM-Agents-Simulation-Framework

control. The foundational 15B-parameter model, Hominis-large, was pretrained on the RedPajama-v2 web corpus [22], extensively filtered for quality and diversity, and further enhanced with curated data, including scientific papers from arXiv and permissively licensed source code from GitHub. This combination supports both broad linguistic competence and domain-specific precision, particularly in technical and factual domains like science, mathematics, and computer science.

Following the pretraining phase, *Hominis-large* underwent extensive instruction tuning to refine its ability to follow human directives, engage in conversational interactions, and execute specific language-based tasks. This stage utilized a combination of established, openly available instruction tuning datasets possessing clear and permissive licensing terms, notably including subsets of the FLAN v2 collection and the Dolly-15k dataset, both of which are licensed under Apache 2.0.

A cornerstone of our research is the introduction of a novel agentic infrastructure designed to govern the inference-time behavior of *Hominis*, moving beyond prevalent paradigms. While much prior work in agentic systems has revolved around the ReAct pattern[23], which interleaves discrete reasoning and action steps in a sequential manner, or more recent multi-execution strategies that permit a greater number of actions per reasoning cycle[24], our methodology diverges significantly. We propose an approach where executable code serves as the primary medium for an agent's actions and interactions with its environment. Instead of relying on loosely structured tool usage through API calls or complex chains of textual prompts, our agent formulates and executes complete programs capable of implementing sophisticated, high-level behaviors such as information retrieval from specified sources, data aggregation from multiple outputs, or even recursive self-correction of its own generated content.

Although the concept of LLM-driven code generation for agency has precedents[25], practical implementations have frequently encountered significant challenges, primarily attributable to the limitations inherent in the choice of programming languages. To surmount these limitations, we formalize a minimalist yet powerful computational substrate, termed the *Agentic VM*. The *Agentic VM* is defined by a restricted set of fundamental programming primitives and well-defined interfaces for system-level interoperability.

The culmination of this design is our framework named *ReAgent*, which empowers the construction of intelligent agents that reason and act through structured program synthesis rather than relying on ad hoc tool invocation or brittle prompt engineering. We further leverage the sophisticated capabilities of the *ReAgent* framework to conduct targeted finetuning of *Hominis-lite*, an 8-billion parameter version of our model, thereby creating regionally specialized variants for Italian and Dutch linguistic contexts.

By enabling the model development process to be partially automated and directed by an agent, we achieve a more efficient pathway to creating highly specialized and culturally attuned language models.

Acknowledgments

This work was partially supported by PNRR MUR Project PE0000013-FAIR. The FAIR project is committed to promoting an advanced vision of Artificial Intelligence, driving research and development in this crucial field and constantly keeping ethical, legal and sustainability considerations in mind

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. Chesney, D. K. Citron, Deepfakes and the new disinformation war, Foreign Affairs (2019).

[2] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021).

[3] T. V. Doan, Z. Wang, N. N. M. Hoang, W. Zhang, Fairness in large language models in three hours, Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (2024).
[4] M. Pikuliak, A. Hrckova, S. Oresko, M. Simko, Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling, arXiv preprint arXiv:2311.18711 (2024).
[5] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, et al., Bbq: A hand-built bias benchmark for question answering, Findings of the Association for Computational Linguistics: ACL 2022 (2022).
[6] U. Gadiraju, et al., "i wouldn't say offensive but...": Disability-centered perspectives on large language models, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (2023).
[7] X. Amatriain, Prompt design and engineering: Introduction and advanced methods, arXiv preprint arXiv:2401.14423 (2024).
[8] K. G. Troitzsch, Social Science Microsimulation, Springer Science & Business Media, 1996.
[9] L. D. K. E. Elliott, Agent-based modeling in the social and behavioral sciences, Nonlinear Dynamics, Psychology, and Life Sciences, Vol. 8, No. 2, April, 2004 (2004).
[10] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, arXiv:2304.03442 (2023).
[11] C. Gao, X. Lan, Z. Lu, J. Mao, J. Piao, H. Wang, D. Jin, Y. Li, S3: Social-network simulation system with large language model-empowered agents, arXiv:2307.14984 (2023).
[12] N. Ghaffarzadegan, A. Majumdar, R. Williams, N. Hosseinichimeh, Epidemic modeling with generative agents, arXiv:2307.04986 (2023).
[13] A. Ferraro, A. Galli, V. La Gatta, M. Postiglione, G. M. Orlando, D. Russo, G. Riccio, A. Romano, V. Moscato, Agent-based modelling meets generative ai in social network simulations, in: International Conference on Advances in Social Networks Analysis and Mining, Springer, 2024, pp. 155–170.
[14] J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds, Science advances 7 (2021) eabf4393.
[15] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, B. Cui, Retrieval-augmented generation for ai-generated content: A survey, arXiv preprint arXiv:2402.19473 (2024).
[16] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, W. Chen, Generation-augmented retrieval for open-domain question answering, arXiv preprint arXiv:2009.08553 (2020).
[17] D. Benfenati, A. M. Rinaldi, C. Russo, C. Tommasino, Gencrawl: A generative multimedia focused crawler for web pages classification, in: Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, INSTICC, SciTePress, 2024, pp. 91–101. doi:10.5220/0012998900003838.
[18] 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, arXiv preprint arXiv:2310.05694 (2023).
[19] Z. Ji, Y. Tiezheng, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating llm hallucination via self reflection, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[20] D. Benfenati, G. M. D. Filippis, A. M. Rinaldi, C. Russo, C. Tommasino, A Retrieval-augmented Generation application for Question-Answering in Nutrigenetics Domain, Procedia Computer Science 246 (2024) 586–595. URL: https://www.sciencedirect.com/science/article/pii/S1877050924025092. doi:https://doi.org/10.1016/j.procs.2024.09.467.
[21] A. M. Rinaldi, A. Romano, C. Russo, C. Tommasino, Fpsrec: Football players scouting recommendation system based on generative ai, in: 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 7141–7150. doi:10.1109/BigData62323.2024.10825692.
[22] M. Weber, D. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, et al., Redpajama: an open dataset for training large language models, Advances in neural information processing systems 37 (2024) 116462–116492.
[23] S. Yao, J. Zhao, D. Li, N. Du, A. Khandelwal, J. Gao, J. Liu, React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629 (2022).

[24] D. Nguyen, V. D. Lai, S. Yoon, R. A. Rossi, H. Zhao, R. Zhang, P. Mathur, N. Lipka, Y. Wang, T. Bui, et al., Dynasaur: Large language agents beyond predefined actions, arXiv preprint arXiv:2411.01747 (2024).

[25] X. Wang, Y. Chen, L. Yuan, Y. Zhang, Y. Li, H. Peng, H. Ji, Executable code actions elicit better llm agents, in: Forty-first International Conference on Machine Learning, 2024.