Interpretable Language Models for Transparent Decision-Making in Business-Critical Domains

Nicolò Mombelli[a]

[a]Dept of Economics and Management, University of Brescia, Italy
ORCID (Nicolò Mombelli): https://orcid.org/0009-0006-0550-7240

Abstract. Large Language Models (LLMs) present significant potential for enhancing decision support in business-critical contexts, while simultaneously posing challenges related to transparency and reliability. Addressing these concerns, in my research I am investigating interpretable and human-in-the-loop frameworks that integrate automated recommendations with natural language justifications. Contributions include the development of interpretable LLM-based systems, alongside novel contrastive explanation techniques for ranking-based decision processes to be applied in corporate environment. Future directions involve advancing mechanistic interpretability, implementing guardrails against adversarial attacks, and broadening the application of LLMs across organizational functions, with the aim of fostering transparent and regulation-compliant AI systems.

1 Introduction

The recent surge in the deployment of LLMs across a variety of applications has brought remarkable advances in natural language processing and automated decision-making [9]. However, as these models increasingly permeate high-stake corporate environments—such as human resources and talent acquisition—the need for transparency and accountability becomes paramount [19]. In such sensitive domains, where algorithmic decisions can significantly impact individuals' careers and organizations' long-term strategies [7], the integration of eXplainable Artificial Intelligence (XAI) and interpretability techniques is no longer optional but essential.

My research activity investigates the integration of LLMs into decision-making workflows across a variety of business domains, where their capacity to process and generate human-like language presents new opportunities for the development of intelligent support systems [20]. The study focuses on understanding how LLMs can be effectively embedded within organizational pipelines to improve decision quality, while ensuring alignment with interpretable and explainable frameworks that meet both regulatory requirements and the needs of end-users. From this perspective, my research contributes to a more comprehensive understanding of LLMs not merely as linguistic technologies, but as strategic assets in the design and implementation of contemporary business decision-making processes.

2 Research Questions

While LLMs exhibit impressive performance across a wide range of task types [4], their integration into decision-support pipelines introduces significant challenges, particularly with respect to trustworthiness, usability, and alignment with human values [10]. Furthermore, the regulatory landscape surrounding these technologies is becoming increasingly stringent, often mandating the incorporation of human-in-the-loop mechanisms to ensure oversight [14] and guardrails strategies to prevent adversarial attacks [8].

To address these concerns, my research studies focus on the following central questions:

1. How can natural language explanations generated by LLMs enhance trust and stakeholder understanding in decision-support systems deployed in business-critical domains?
2. What design principles enable the effective integration of language models with interpretability techniques to deliver reliable decision-making tools?
3. How can human-in-the-loop methodologies and appropriate evaluation metrics be employed to assess the quality, consistency, and domain alignment of automated recommendation systems?

Through addressing these research questions, the goal of my studies is to contribute to the growing body of literature at the intersection of LLM deployment, interpretability, and augmented decision-making, while grounding the investigation in applied, real-world corporate settings.

3 Contributions

In response to theses proposed research questions, my studies to date have produced a set of methodological and applied contributions across three main axes.

LLM Framework for Transparent Recruitment In collaboration with a major European banking institution—where I serve as a Senior Data Scientist—I designed and empirically validated an operational framework for AI-assisted résumé evaluation [13]. The system integrates quantitative candidate profiling with natural language explanations generated by LLMs, offering recruiters fine-grained, interpretable assessments of candidates' curricula vitae. This approach is explicitly designed to enhance transparency and support regulatory compliance, particularly in alignment with the provisions of the latest European AI Act [5]. Importantly, the role of LLMs in this framework is not to replace human resource professionals, but rather to augment decision-making by supporting recruiters with evidence-based guidance and structured inputs for candidate interviews. A human-in-the-loop protocol is implemented, allowing recruiters to

review and provide feedback on both the LLM-generated scores and explanations, while ultimately retaining control over the evaluation and selection process. The evaluative framework includes stability testing for the generated explanations and incorporates a structured validation methodology to assess the alignment of LLM outputs with expert-defined gold standards. Additionally, the model's ability to extract relevant and factually accurate information from résumés is assessed. This work makes a concrete contribution to the development of explainability-aware AI systems tailored to high-risk domains, such as hiring, where trust, interpretability, and regulatory accountability are paramount.

XAI to Support Decision-Making in Corporate Banking Expanding upon the need for stakeholder-aligned explainability, I extended the application of LLMs to generate business-relevant justifications for AI-driven recommendations in the corporate banking sector [3]. Specifically, this study focused on supporting sales personnel by not only providing propensity-to-buy scores, but also delivering natural language explanations that articulate the underlying rationale behind each recommendation. The proposed methodology involves clustering input features based on their business semantics and leveraging LLMs to translate their associated SHAP values into coherent, human-readable narratives. The resulting explanations were benchmarked against expert-generated justifications, demonstrating both alignment with domain expectations and measurable improvements in operational efficiency. This approach enabled the deployment of a system capable of generating tailored textual justifications for non-technical end-users, such as relationship managers. As a result, the system facilitates scalable, intelligible, and contextually meaningful communication of model outputs, thereby enhancing the transparency and usability of AI recommendations in commercial banking workflows.

Contrastive Explanation Techniques for Ranking Systems To enhance the interpretability of ranking-based decision-making systems, I introduced the concept of Evaluative Item-Contrastive Explanations as a novel form of contrastive reasoning [2]. This approach is based on the pairwise comparison of items ranked by an AI model, whereby a comprehensive evaluative assessment is conducted for each item in the pair. Specifically, the method generates natural language representations that articulate the strengths and weaknesses of both the preferred and the non-selected item, thereby providing users with a nuanced understanding of the rationale behind the ranking outcome. Grounded in the principles of Evaluative AI [12], this technique does not merely justify the selection of the top-ranked candidate but also highlights the distinguishing characteristics of alternatives, fostering a more transparent and accountable decision-making process. The method holds particular promise for high-stakes domains such as recruitment, where fairness and clarity are critical, as well as for resource-constrained environments that demand optimized decisions under strict budgetary limitations.

Collectively, these contributions advance the state of the art in decision-making and XAI by demonstrating how interpretability techniques—particularly in conjunction with LLMs—can be meaningfully applied and evaluated in real-world environments. The research highlights both technical innovations and socio-technical implications, laying the foundation for future inquiry into trustworthy and human-aligned AI systems.

4 Future Research

Building upon the foundational contributions outlined above, future research will pursue three intertwined directions: (i) deepening the interpretability of LLM-based systems through both user-centric and mechanistic perspectives, (ii) developing robust guardrails to mitigate adversarial attacks against such systems, and (iii) advancing the integration of language models into decision-making pipelines across business domains.

First, on the interpretability front, while prior efforts have emphasized natural language explanations and human-aligned justification strategies, future work will investigate the internal representations and reasoning processes of LLMs themselves [11]. Drawing inspiration from recent advances in mechanistic interpretability [18], my research will explore how internal attention patterns, neuron activations, and intermediate representations can be probed to reveal the computational substrates of decisions made by LLMs. The goal is to bridge the gap between surface-level explainability (what the model says) and structural transparency (how the model reasons), thereby contributing to a more principled understanding of LLM behavior in complex decision contexts.

Second, as these LLM-based systems are increasingly integrated into high-stakes corporate environments, they become potential targets for manipulation, data poisoning, or prompt-based exploits that can compromise both reliability and trust [8]. By designing effective guardrailing mechanisms, my goal is to ensure that automated recommendations remain aligned with organizational objectives, resilient under adversarial conditions, and compliant with regulatory and ethical standards [16].

Finally, from an application-driven standpoint, future work will focus on extending LLM-based decision-support frameworks to additional high-impact business functions, such as performance evaluation and client risk assessment. Particular attention will be paid to the longitudinal impact of AI-generated explanations on organizational processes, decision quality, and stakeholder trust [1], with an emphasis on designing adaptive systems that evolve with business rules and user feedback [15].

Additionally, future research could address the development of evaluation protocols that jointly assess explanation quality and factual accuracy [21], along with the alignment with expert judgment [17]. These metrics could be validated not only via human-in-the-loop processes but also through the use of synthetic benchmarks and automated auditing tools [6], with the aim of supporting scalable deployment in regulated environments.

References

[1] M. Amirizaniani, J. Yao, A. Lavergne, E. S. Okada, A. Chadha, T. Roosta, and C. Shah. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*, 2024.

[2] A. Castelnovo, R. Crupi, N. Mombelli, G. Nanino, and D. Regoli. Evaluative item-contrastive explanations in rankings. *Cognitive Computation*, 16(6):3035–3050, 2024.

[3] A. Castelnovo, R. Depalmas, F. Mercorio, N. Mombelli, D. Potertì, A. Serino, A. Seveso, S. Sorrentino, and L. Viola. Augmenting xai with llms: A case study in banking marketing recommendation. In *World Conference on Explainable Artificial Intelligence*, pages 211–229. Springer, 2024.

[4] D. Cheng, S. Huang, and F. Wei. Adapting large language models to domains via reading comprehension. *arXiv preprint arXiv:2309.09530*, 2023.

[5] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and

amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, 6 2024. Official Journal of the European Union.

[6] P. Hämäläinen, M. Tavast, and A. Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[7] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[8] A. Kumar, C. Agarwal, S. Srinivas, A. J. Li, S. Feizi, and H. Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

[9] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyürek, A. Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.

[10] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[11] S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

[12] T. Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 333–342, 2023.

[13] N. Mombelli, R. Crupi, S. Rubini, A. Ermellino, N. Alborè, F. Rufini, G. Cosentini, Andrea Claudio Greco, and A. Castelnovo. Evaluating ai-generated explanations: A case study on llm-assisted candidate assessmen, 2025.

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, and P. Christiano. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 35 (NEURIPS 2022)*, Advances in Neural Information Processing Systems, 2022. ISBN 978-1-7138-7108-8. 36th Conference on Neural Information Processing Systems (NeurIPS), ELECTR NETWORK, NOV 28-DEC 09, 2022.

[15] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

[16] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

[17] P. D. Tailor, L. A. Dalvin, M. R. Starr, D. A. Tajfirouz, K. D. Chodnicki, M. C. Brodsky, S. A. Mansukhani, H. E. Moss, K. E. Lai, M. W. Ko, et al. A comparative study of large language models, human experts, and expert-edited large language models to neuro-ophthalmology questions. *Journal of Neuro-Ophthalmology*, pages 10–1097, 2022.

[18] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, A. Tamkin, E. Durmus, T. Hume, F. Mosconi, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024.

[19] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[20] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

[21] A. Yalamanchili, B. Sengupta, J. Song, S. Lim, T. O. Thomas, B. B. Mittal, M. E. Abazeed, and P. T. Teo. Quality of large language model responses to radiation oncology patient care questions. *JAMA network open*, 7(4):e244630–e244630, 2024.