Chapter 4
Are LLMs and RAG Trustworthy Enough for Your Business?
A Deep Dive Into AI's Reliability

Hakan Emekci
https://orcid.org/0000-0002-4074-5600
TED University, Turkey

ABSTRACT

Integrating Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems into business operations has become a transformative force. In the literature, the advantages as well as the disadvantages of this integration are discussed. The chapter discusses some latent risks involved in AI, namely, AI-data bias, adversarial vulnerabilities, privacy concerns, and domain-specific limitations, while putting forward methodologies for their mitigation through better data management and strong security protocols. The chapter also discusses the prospect of the future of changes that could be made in reliable and safe AI. Stitching together these divergent insights, this paper contributes to an understanding of how businesses can use LLMs and RAG systems responsibly in ways that keep AI adoption in step with meeting ethical compasses and operational integrity. It forms a core reference to guide institutions in their efforts to harness the full potential of artificial intelligence technologies in their operations regarding the intricacies involved in trust and risk management.

1. INTRODUCTION

Throughout the last couple of years, fast technology integration appeared in association with AI, mainly linked to the appearance of large language models. This transformed many business environments in different industries (Al Ghadban et al., 2023; Chen et al., 2023). These highly advanced AI models are widely acknowledged as one of the key tools in boosting efficiency, enabling real-time decision-making, and offering personalized customer experiences, ranging from automation in customer support to high-level strategic data analysis (Asai et al., 2023; Jeong, 2023). However, as AI models, such as LLMs and RAG, become integral to core business functions, the need for their reliability to be scrutinized intensifies. Indeed, while these models promise high accuracy and adaptability, their trustworthiness remains one of the most important areas of concern, as businesses attempt to strike a balance between

DOI: 10.4018/979-8-3693-8332-2.ch004

innovating with new capabilities and managing risks and compliance (Gámiz et al., 2021; Payette & Abdul-Nour, 2023). These would include metrics such as accuracy, robustness, transparency, and ethical compliance. Each of these becomes an important dimension in establishing the business viability of LLMs and RAG. For example, the accuracy in model predictions is a direct determinant of the reliability of automated decisions, while robustness speaks to a model's ability to resist unpredicted inputs or perturbations (Muneeswaran et al., 2023). This transparency is one avenue to satisfying not just regulatory constraints but also a user's quest to understand and, hence, trust the model's decisions. These decisions may vary depending on the level of interpretability embedded in the architecture of the AI (Linardatos et al., 2021). Ethical and legal considerations underline the requirement for equity in decision-making, with sensitive domains including healthcare, by Dale et al. (2023), and finance, as discussed by Cao (2022), since the models' outputs might affect individual lives and have serious wider consequences in society. The main problem is that they are big datasets, some of which may be crawled into diverse, uncontrolled environments.

This dependency also exposes them to risks relating to data bias, which may skew decision-making processes. Many AI fairness studies have illustrated that when models are trained on imbalanced datasets, they tend to further propagate the already existing biases and thus affect the ethical integrity of AI-driven decisions (Navigli et al., 2023). More importantly, LLMs and RAG systems are vulnerable to adversarial attacks; maliciously designed inputs can manipulate the model's output (Shayegani et al., 2023). The existence of such vulnerabilities is especially alarming for enterprises that manage sensitive information, since these attacks not only undermine the reliability of models but may also result in insubstantial data breaches and harm to reputation. The ability of artificial intelligence systems to impact business operations necessitates rigorous security protocols and extensive lifecycle management (Raimundo & Rosário, 2021).

It requires, therefore, great vigilance and frequent updates so that the model remains able to combat the new threats and adapt to the changes in situations. These models can be further bolstered with the latest adversarial training methods against their possible manipulations, making them more reliable in realistic application scenarios (Zhao et al., 2022). In this aspect, the governance of enterprise and industry-specific adaptability of needs for AI models turn out to be an important differentiator in deploying trustworthy AI.

This chapter aims to critically analyze the application of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems as they apply to corporate environments, determining both the strategic advantages that they offer and the risks involved when implemented. The investigation starts by exploring the differences between the foundational designs of LLM and RAG systems and their potential use within business contexts. Then, the chapter compares these technologies against key criteria, including accuracy, dependability, transparency, ethical standards, and compatibility with regulatory systems. The methodology includes an extensive review of current research literature, analysis of key case studies, and investigation of current trends of application. Through the synthesis of findings, this research presents an organizational strategy to apply LLM and RAG technologies effectively and ethically.