

Risk Identification and Management in Hybrid Cognitive Systems Based on Large Language Models

Edervaldo Jose de Souza Melo
Independent Researcher
Campo Grande, MS, Brazil
ORCID: 0009-0003-6835-135X
email: edersouzamelo@gmail.com

Abstract—This paper presents an exploratory technical report on risk identification and management in hybrid cognitive systems based on large language models (LLMs). Rather than focusing on system performance or benchmark evaluation, the study concentrates on non-functional risks related to epistemic uncertainty, operational misuse, architectural opacity, and governance constraints. The objective is to document a structured descriptive framework for anticipating potential failure modes in early-stage cognitive architectures, with emphasis on traceability and analytical clarity prior to full system implementation.

Index Terms—Hybrid cognitive systems, risk management, large language models, AI governance, non-functional requirements

I. INTRODUCTION

Recent advances in large language models have enabled the emergence of hybrid cognitive systems that combine symbolic structures, procedural rules, and probabilistic inference with generative language capabilities. While these systems exhibit promising flexibility, they also introduce new categories of risk that are not adequately addressed by traditional software engineering practices. This paper argues that early-stage risk identification constitutes a relevant analytical step prior to performance evaluation or large-scale deployment.

II. SCOPE AND METHODOLOGICAL POSITIONING

This work adopts an exploratory and qualitative methodological stance. It does not claim empirical validation, benchmark superiority, or statistical generalization. Instead, it positions itself as a technical mapping exercise aimed at identifying risk categories and failure modes observable during the conceptual and pre-implementation phases of hybrid cognitive system design.

III. RISK TAXONOMY FOR HYBRID COGNITIVE SYSTEMS

Based on architectural analysis and iterative design reflection, four primary classes of non-functional risk are identified: epistemic risks related to model uncertainty and interpretability; operational risks associated with misuse or misalignment in real-world contexts; architectural risks arising from system opacity and component coupling; and governance risks linked to accountability, responsibility allocation, and lifecycle control.

IV. RISK IDENTIFICATION AND MAPPING PROCEDURE

The proposed procedure involves iterative risk elicitation aligned with system design stages. Risks are documented through structured descriptors, including origin, affected components, potential impact, and mitigation constraints. This approach emphasizes traceability and revision rather than premature optimization or quantitative scoring.

V. DISCUSSION AND LIMITATIONS

The absence of empirical benchmarks and implementation artifacts limits the scope of the present analysis. However, this limitation is intentional, as the contribution of this paper lies in documenting a disciplined risk-oriented mindset applicable before system stabilization. Future work may integrate this taxonomy with empirical validation once implementation data becomes available.

VI. CONCLUSION AND FUTURE WORK

This paper contributes a structured perspective on risk management for hybrid cognitive systems based on LLMs, emphasizing non-functional considerations often neglected in early design phases. Future research directions include empirical validation, integration with software lifecycle models, and alignment with emerging AI governance standards.

REFERENCES

- [1] I. Sommerville, *Software Engineering*, 10th ed. Boston, MA, USA: Pearson, 2016.
- [2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Upper Saddle River, NJ, USA: Pearson, 2021.
- [3] B. Mittelstadt et al., "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, 2016.
- [4] ISO/IEC 25010, "Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE)," 2011.
- [5] ISO/IEC 23894, "Artificial intelligence – Risk management," 2023.