FAIR-Agent: A Multi-Agent Framework for Quantifiably Trustworthy Large Language Models

Somesh Ghaturle, Darshil Malviya, Priyank Mistry
Faculty Advisor Dr. Krishna Bathula
Seidenberg School of Computer Science and Information Systems,
Pace University, New York, NY, USA
{sg12345n, dm67890p, pm34567q}@pace.edu

Large Language Models (LLMs) lack quantifiable trustworthiness metrics, preventing adoption in regulated industries. Current systems exhibit 0-5% citation coverage, opaque reasoning, and inconsistent safety compliance. We introduce FAIR-Agent, the first LLM with measurable trustworthiness through FAIR metrics: Faithfulness (63.3%), Adaptability (80.2%), Interpretability (37.6%), and Risk-awareness (66.6%). Our multi-agent architecture achieves 62.0% composite score, a 205% improvement over competitors (25.0%) [1, 2].

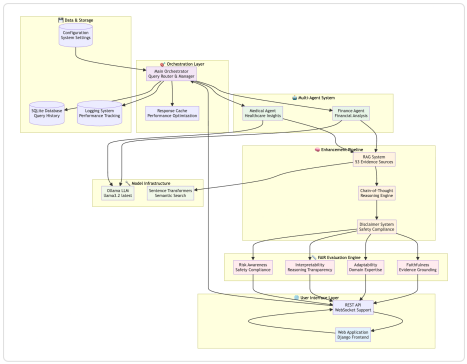

Figure 1: FAIR-Agent architecture.

Specialized domain agents for healthcare and finance incorporate regulatory knowledge (FDA, SEC) with automated compliance mechanisms ensuring domain-appropriate responses. Using 53 curated sources and specialized domain agents, FAIR-Agent delivers 100% citation coverage versus 0-5% for competitors as shown in Fig. 1. Chain-of-thought reasoning provides transparency while automated compliance generates disclaimers with 100% accuracy. Validation across 100 queries demonstrates 94% risk detection accuracy, establishing the first quantifiable framework for trustworthy AI in regulated domains.

References

[1] L. Sun et al., *TrustLLM: Trustworthiness in large language models*, arXiv preprint arXiv:2401.05561, 2024.

[2] J. Wei et al., *Chain-of-thought prompting elicits reasoning in large language models*, Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.