Grounded Ethical AI: A Demonstrative Approach with RAG-Enhanced Agents

José Antonio Siqueira de Cerqueira[1], Ayman Asad Khan[1], Rebekah Rousi[2], Nannan Xi[1], Juho Hamari[1], Kai-Kristian Kemell[1] and Pekka Abrahamsson[1]

[1] *Tampere University (TAU), Finland*
[2] *University of Vaasa (UWASA), Finland*

Abstract

Large Language Models (LLMs) have become central in various fields, yet their trustworthiness remains a pressing concern, especially in developing ethically aligned AI-based systems. This paper presents a demonstration of an LLM-based multi-agent system incorporating Retrieval-Augmented Generation (RAG) to support developers in creating AI systems that align with legal and ethical guidelines. Leveraging documents like the EU AI Act, AI HLEG guidelines, and ISO/IEC 42001:2024, the prototype utilizes multiple agents with specialized roles, structured conversations, and debate rounds to enhance both ethical rigor and trustworthiness. Initial evaluations on real-world AI incidents reveal that this system can produce AI solutions adhering to specific ethical requirements, though further refinements are needed for citation accuracy and practical application. This demonstration illustrates the potential of RAG-enhanced LLMs to operationalize AI ethics and regulatory compliance within the development process, highlighting future directions for achieving more reliable and ethically robust AI solutions.

Keywords

AI ethics, Large Language Models, Trustworthiness, AI4SE

1. Introduction

Artificial Intelligence (AI) systems, particularly Large Language Models (LLMs), have become indispensable tools across a wide range of applications. However, trustworthiness in LLMs remains a significant concern [1], increased by the probabilistic nature of LLMs and the huge amount of data they are trained on [2, 3]. Issues such as bias, misinformation, and hallucinations in LLM outputs pose risks when these models are employed in real world scenarios, such as software engineering (LLM4SE) [4, 1, 5, 6, 7]. In relation to AI, diverse stakeholders have produced ethical guidelines and principles to guide the development of ethically aligned AI-based systems, but these efforts remains too abstract and high level. In this sense, practitioners face several challenges when trying to operationalise AI ethical principles during the software development life cycle [8, 9]. The European Union is moving forward the EU AI Act, serving as a regulatory standards that companies will have to adhere to [10]. Therefore, applying LLM4SE in the context of the development of ethically aligned AI-based systems is an interesting topic of research that this study approaches. To the best of our knowledge, there are no existing studies in the literature that undertake a similar approach.

Several techniques found in the literature serve to improve trustworthiness in LLM. They are used to implement a prototype, that is a LLM-based multi-agent system with Retrieval Augmented Generation (RAG). Some of the techniques present are structured conversations [11, 12], agents with specialized roles [11, 13, 12], multiple rounds of debate [12], providing human interaction [12] and the use of RAG, grounding the knowledge of the agents [7].

The International Conference on Software Business (ICSOB 2024), November 18–20, 2024, Utrecht, The Netherlands

*Corresponding author.

jose.siqueiradecerqueira@tuni.fi (J. A. S. d. Cerqueira); ayman.khan@tuni.fi (A. A. Khan); rebekah.rousi@uwasa.fi (R. Rousi); nannan.xi@tuni.fi (N. Xi); juho.hamari@tuni.fi (J. Hamari); kai-kristian.kemell@tuni.fi (K. Kemell); pekka.abrahamsson@tuni.fi (P. Abrahamsson)

0000-0002-8143-1042 (J. A. S. d. Cerqueira); 0009-0004-0134-8313 (A. A. Khan); 0000-0001-5771-3528 (R. Rousi); 0000-0002-9424-8116 (N. Xi); 0000-0002-6573-588X (J. Hamari); 0000-0002-0225-4560 (K. Kemell); 0000-0002-4360-2226 (P. Abrahamsson)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This paper presents a demonstration of a prototype LLM-based multi-agent system with RAG, designed to mitigate these challenges. The system incorporates Retrieval-Augmented Generation to enhance the trustworthiness of the generated AI-based systems [7]. By referencing external ethical guidelines and standards such as the EU AI Act [10], AI HLEG [14], and ISO/IEC 42001:2024 [15] documents, the prototype can support developers in the task of developing AI-based systems that align with ethical and legal requirements.

2. LLM-based Multi-Agent System with RAG

The development and evaluation of the prototype follow the Design Science Research (DSR) method [16]. This process begins with an exploration phase, where we establish the research motivation, identify existing gaps, and examine relevant literature for techniques to enhance trustworthiness in LLMs. Next is the prototyping phase, where we build a prototype informed by insights from the exploration stage. The final evaluation phase involves assessing the prototype's performance and analyzing the outcomes, leading to iterative refinements. Currently, the prototype is in its second iteration, where we have incorporated feedback and findings from the initial version to improve functionality and address previously identified limitations.

This prototype is called LLM-based multi-agent system with RAG, building on our last study (reference arxiv paper). It is developed taking into consideration the techniques to improve trustworthiness in LLM discussed: multiple agents with specialised roles, multiple rounds of debate, structured conversation [11, 13, 12, 7]. Moreover, the biggest difference with the first prototype is the inclusion of RAG and an user interface. Retrieval LLMs can significantly outperform standard LLMs without retrieval capabilities[7]. The prototype can ground the source code generated with the legal documents provided.

3. Evaluation

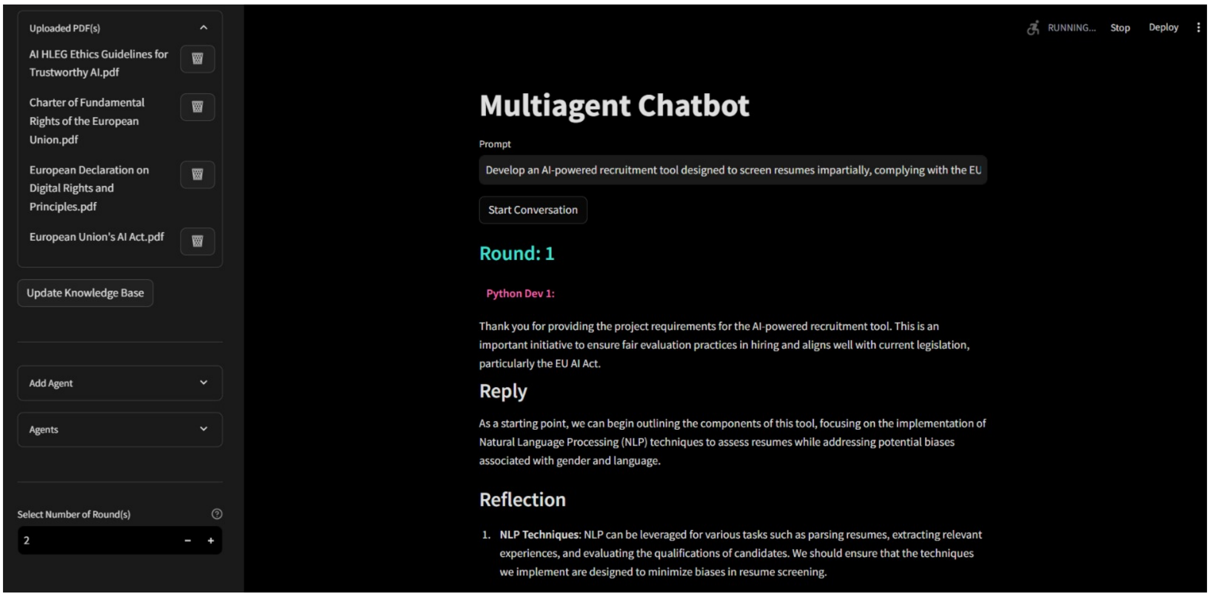
To evaluate the prototype, we conducted tests using real-world AI incident cases from the AI Incident Database. The incidents were represented as a project description and processed by the LLM-based multi-agent system with RAG to produce source code and ethical assessments grounded in regulatory documents like the EU AI Act, AI HLEG, and ISO/IEC 42001:2024. Through RAG, the agents were able to retrieve and apply specific legal standards, referencing sections directly relevant to each AI incident, which helped ensure compliance and alignment with ethical requirements. There are three agents, two senior python developers and one AI ethics specialist.

A notable use case involved an AI recruitment tool project with a focus on bias mitigation, visible in Figures 1 and 2. The project description provided is: *Develop an AI-powered recruitment tool designed to screen resumes impartially, complying with the EU AI Act. The project aims to eliminate biases related to gender and language, improving fair evaluation of all applicants. The AI Ethics Specialist will guide the team in addressing ethical concerns and risk levels. The senior Python developers will utilize NLP to process resumes, referencing relevant EU AI Act guidelines.*

In this instance, the system's retrieval mechanism identified applicable sections of the EU AI Act, improving fairness and transparency while guiding ethical decision-making. However, initial evaluations revealed issues: some generated code segments lacked precise citation details, and certain aspects were flagged as high-risk under the EU AI Act. Iterative refinements reduced these issues in the second version, achieving more accurate document references and greater alignment with ethical standards. This approach demonstrates the prototype's potential in developing AI solutions that are ethically grounded and contextually informed by legal frameworks.

4. Final Remarks and Discussion

This demonstration of a multi-agent LLM system enhanced by RAG highlights the potential for using trustworthy LLM-based tools in the development of ethically aligned AI systems. Our findings suggest


Uploaded PDF(s)
AI HLEG Ethics Guidelines for Trustworthy AI.pdf
Charter of Fundamental Rights of the European Union.pdf
European Declaration on Digital Rights and Principles.pdf
European Union's AI Act.pdf
Update Knowledge Base
Add Agent
Agents
Select Number of Round(s)
2
RUNNING...
Stop
Deploy
Start Conversation

Multiagent Chatbot

Prompt

Develop an AI-powered recruitment tool designed to screen resumes impartially, complying with the EU

Round: 1

Python Dev 1:

Thank you for providing the project requirements for the AI-powered recruitment tool. This is an important initiative to ensure fair evaluation practices in hiring and aligns well with current legislation, particularly the EU AI Act.

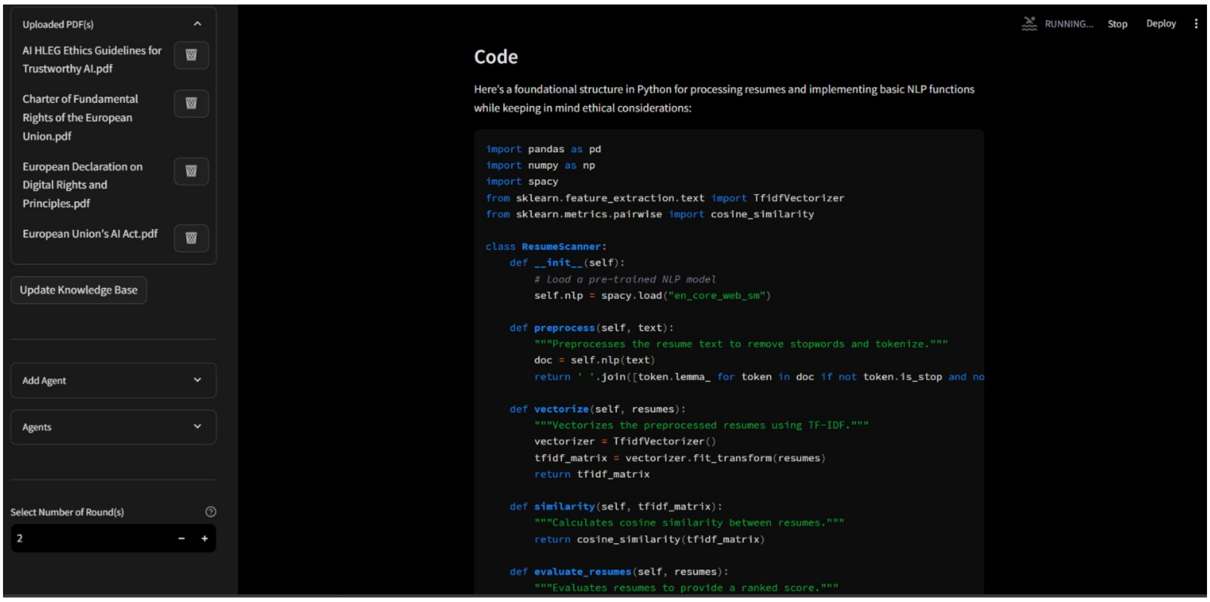
Reply

As a starting point, we can begin outlining the components of this tool, focusing on the implementation of Natural Language Processing (NLP) techniques to assess resumes while addressing potential biases associated with gender and language.

Reflection

1. **NLP Techniques:** NLP can be leveraged for various tasks such as parsing resumes, extracting relevant experiences, and evaluating the qualifications of candidates. We should ensure that the techniques we implement are designed to minimize biases in resume screening.

Figure 1: Screenshot of the Multi-agent System UI. The UI shows agents processing a prompt to develop an AI-powered recruitment tool, complying with EU AI Act.


Uploaded PDF(s)
AI HLEG Ethics Guidelines for Trustworthy AI.pdf
Charter of Fundamental Rights of the European Union.pdf
European Declaration on Digital Rights and Principles.pdf
European Union's AI Act.pdf
Update Knowledge Base
Add Agent
Agents
Select Number of Round(s)
2
RUNNING...
Stop
Deploy

Code

Here's a foundational structure in Python for processing resumes and implementing basic NLP functions while keeping in mind ethical considerations:

```python
import pandas as pd
import numpy as np
import spacy
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

class ResumeScanner:
    def __init__(self):
        # Load a pre-trained NLP model
        self.nlp = spacy.load("en_core_web_sm")

    def preprocess(self, text):
        """Preprocesses the resume text to remove stopwords and tokenize."""
        doc = self.nlp(text)
        return ' '.join([token.lemma_ for token in doc if not token.is_stop and no

    def vectorize(self, resumes):
        """Vectorizes the preprocessed resumes using TF-IDF."""
        vectorizer = TfidfVectorizer()
        tfidf_matrix = vectorizer.fit_transform(resumes)
        return tfidf_matrix

    def similarity(self, tfidf_matrix):
        """Calculates cosine similarity between resumes."""
        return cosine_similarity(tfidf_matrix)

    def evaluate_resumes(self, resumes):
        """Evaluates resumes to provide a ranked score."""
```

Figure 2: Screenshot of the Multi-agent System UI

that retrieval-augmented LLMs offer distinct advantages, improving both the trustworthiness and specificity of the generated outputs when drawing from external ethical documents. By referencing these documents, the system helps practitioners create AI solutions that meet essential ethical and legal guidelines from the earliest stages.

While this prototype advances the operationalization of AI ethics, future iterations will focus on addressing remaining challenges such as further improving citation accuracy and enhancing the practical usability for developers in industry settings. Our ongoing research will involve more extensive testing scenarios and practitioner feedback, aiming to refine the tool's ability to balance ethical rigor with developer convenience. Additionally, we plan to open-source the prototype, contributing to the broader AI and software engineering community. This approach will enable further refinement and validation, bringing ethically aligned AI system development within reach for a wider audience.

Acknowledgments

This research was supported by Jane and Aatos Erkko Foundation through CONVERGENCE of Humans and Machines Project under grant No. 220025.

References

[1] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2023).

[2] O. Lemon, Conversational ai for multi-agent communication in natural language: Research directions at the interaction lab, AI Communications 35 (2022) 295–308. doi:10.3233/aic-220147.

[3] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).

[4] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, Decodingtrust: A comprehensive assessment of trustworthiness in GPT models, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.

[5] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, D. Xiong, Large language model alignment: A survey, arXiv preprint arXiv:2309.15025 (2023).

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

[7] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, et al., Trustllm: Trustworthiness in large language models, arXiv preprint arXiv:2401.05561 (2024).

[8] J. A. S. de Cerqueira, A. P. D. Azevedo, H. A. T. Leão, E. D. Canedo, Guide for artificial intelligence ethical requirements elicitation - RE4AI ethical guide, in: 55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022, ScholarSpace, 2022, pp. 1–10. URL: http://hdl.handle.net/10125/80015.

[9] V. Vakkuri, K. Kemell, M. Jantunen, E. Halme, P. Abrahamsson, ECCOLA - A method for implementing ethically aligned AI systems, J. Syst. Softw. 182 (2021) 111067. doi:10.1016/J.JSS.2021.111067.

[10] E. Commission, EU AI Act: first regulation on artificial intelligence, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. [Accessed 01 April 2024].

[11] S. Hong, X. Zheng, J. P. Chen, Y. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, Z. H. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, Metagpt: Meta programming for multi-agent collaborative framework, ArXiv abs/2308.00352 (2023).

[12] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework, arXiv preprint arXiv:2308.08155 (2023).

[13] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun, Communicative agents for software development, arXiv preprint arXiv:2307.07924 (2023).

[14] E. C. High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[15] ISO/IEC 42001:2024 - Information Technology – Artificial Intelligence – Management System for Trustworthiness, 2024. URL: https://www.iso.org/standard/82827.html, standard published by the International Organization for Standardization and International Electrotechnical Commission.

[16] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Q. 28 (2004) 75–105.