TrustInLLM

Trustworthy Digital Systems Assisted by Large Language Models (LLMs)

Programm / Ausschreibung	Digitale Technologien, Digitale Technologien, Digitale Schlüsseltechnologien: Ausschreibung 2023	**Status**	laufend
Projektstart	01.10.2024	**Projektende**	30.09.2027
Zeitraum	2024 - 2027	**Projektlaufzeit**	36 Monate
Keywords	Trust-related Requirements Engineering, Large Language Models (LLMs), Safety-critical Systems, AI Act		

Projektbeschreibung

Die europäische Gesellschaft verlässt sich zunehmend auf komplexe digitale Systeme. In diesem Zusammenhang adressieren wir das Problem, effizient und zuverlässig unter Verwendung einer großen und komplexen Menge von miteinander verknüpften technischen Dokumenten von digitalen Systemen zu arbeiten, bei denen selbst menschliche Experten kaum den Überblick behalten oder alle Details kennen können. Dies ist besonders wichtig in sicherheitskritischen Bereichen wie etwa dem von Elektrofahrzeugen, in denen solche Dokumentation für das korrekte Spezifizieren von Hardware-Software-Schnittstellen erforderlich ist. Traditionelle Methoden der Informationssuche ermöglichen keinen einfachen Zugriff auf die Informationen in solchen Dokumenten, da sie Expertenkenntnisse in der Verwendung spezifischer Abfragesprachen erfordern, das z.B. Safetyexperten meist nicht haben.
Aktuelle Fortschritte in der Generativen Künstlichen Intelligenz (GenAI) und insbesondere in Large Language Models (LLMs) legen nahe, dass für das vorliegende Problem Dialoge in natürlicher Sprache möglich sein werden, um auf den Inhalt einer solchen Dokumentensammlung zuzugreifen. Beim aktuellen Stand der Technik besteht jedoch das Hauptproblem der „Halluzinationen“ von LLMs, d.h. das Antworten mit Informationen, die nicht auf faktischem Wissen basieren, sondern eher aus „kreativen“ Generierungen durch das Sprachmodell stammen.
Dies ist ein wesentliches Hindernis bezüglich des Vertrauens in solche Antworten und im Allgemeinen für einen solchen Ansatz, insbesondere wenn er in einem sicherheitskritischen Bereich wie Elektrofahrzeugen angewendet werden könnte. Das Vermeiden solcher Halluzinationen ist entscheidend, um Sicherheitsrisiken bei der Entwicklung solcher Fahrzeuge zu reduzieren.
Vertrauen ist ein komplexes psychologisches und soziologisches Konzept, und die Vertrauenswürdigkeit eines potenziell Vertrauenswürdigen ist ein wesentlicher Aspekt bei der Etablierung von Vertrauen. Die Bedeutung von Vertrauen beschränkt sich nicht nur auf den zwischenmenschlichen Bereich; Vertrauen kann auch die Art und Weise definieren, wie Menschen mit Technologie interagieren. Das Vertrauen in Automatisierung wurde von Vertrauensbeziehungen zwischen Menschen übertragen. Der Vertrauende ist der Mensch, der potenziell Vertrauenswürdige ist ein automatisiertes System.
Daher adressieren wir die Gewährleistung von Vertrauen durch die Schaffung eines durch LLM-unterstützten Ansatzes, der vertrauenswürdig ist. Dies steht im Einklang mit der Verordnung über künstliche Intelligenz (AI Act), die gerade vom Europäischen Parlament verabschiedet wurde. Der AI Act besagt:

"The purpose of this Regulation is to promote the uptake of human centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law and the environment from harmful effects of artificial intelligence systems in the Union while supporting innovation and improving the functioning of the internal market. ...".
Da der AI Act ein rechtlicher Rahmen ist, umfasst unser Konsortium auch Rechtsexperten, die sicherstellen werden, dass unsere Forschung und Entwicklung tatsächlich im Einklang mit dem AI Act stehen und dessen Anwendung unterstützen. Um unseren Ansatz menschenzentriert zu gestalten, umfasst unser Konsortium auch eine Expertin für menschenzentriertes Design, die auf Diversity und Geschlechtergerechtigkeit achtet. In unserem vorgeschlagenen Projekt beinhaltet dies die Forschung zu einem neuen Ansatz für die Anforderungsanalyse zur Spezifizierung der Bedürfnisse in diesem Zusammenhang.
Unsere vorgeschlagene Forschung zu vertrauensbezogenem Requirements Engineering (RE) wird höchst innovativ sein, indem sie Qualitätsattribute (im Sinne von Eigenschaften) sowohl von Systemen (in unserem Fall LLM-unterstützte Systeme für Abfragen zu technischer Dokumentation) als auch von Daten (in unserem Fall hauptsächlich komplexe technische Dokumentation) miteinander integriert, da Eigenschaften sowohl von Systemen (wie Transparenz) als auch von Daten (wie Korrektheit) relevant sind. Beim aktuellen Stand der Technik werden nur Qualitätsattribute von Systemen in der RE berücksichtigt, aber die Bedeutung von Daten und deren Qualitäten nimmt ständig zu. Daher sind für jedes nicht-triviale datenbasierte System sowohl die Qualitätsattribute des Systems als auch seiner Daten wesentlich, und unser neuer RE-Ansatz wird dies bei der Entwicklung vertrauenswürdiger Systeme berücksichtigen, insbesondere unterstützt durch LLMs. Basierend auf konkreten Anforderungen an ein spezifisches durch LLM unterstütztes System, die gemäß diesem neuen Ansatz definiert werden, wird ein spezifisches Tool, das einem wissenschaftlichen Partner in diesem Konsortium gehört (AERIALL, das freie Open-Source-Modelle wie Mistral unterstützt), weiterentwickelt, um diese Anforderungen zu erfüllen und es vertrauenswürdiger zu machen. Besonderes Augenmerk wird daraufgelegt, Halluzinationen zu vermeiden, unter Berücksichtigung von Eigenschaften sowohl des Systems als auch seiner Daten.
Dieser Ansatz wird mit realen Dokumenten evaluiert werden, die dem industriellen Partner in diesem Konsortium gehören. Zu gegebener Zeit wird er mit dem bekannten ChatGPT-Ansatz von OpenAI verglichen werden, der möglicherweise sogar weniger leistungsfähig ist als unser vorgeschlagener Ansatz für benutzerdefinierte Informationen (in unserem Fall große und komplexe Dokumentation von Hardware-Software-Schnittstellen in einem sicherheitskritischen Bereich) und der insbesondere nicht vertrauenswürdig ist.
Mit diesen Innovationen wird das Konsortium einen bedeutenden Beitrag zur Anwendung modernster KI-Technologie auf ein wichtiges reales Problem leisten, unter Berücksichtigung von sicherheitskritischen Anwendungen. Unser Gesamtansatz ist neu und daher bisher weder in Österreich noch in Europa oder im Ausland verfügbar. Er wird nahtlos den AI Act mit einem neuen Ansatz für Requirements Engineering für vertrauenswürdige Systeme verknüpfen und dessen Einsatz für einen durch LLM unterstützten Ansatz zur Arbeit mit einer großen und komplexen Menge an technischer Dokumentation ermöglichen. Dies ist besonders wichtig für sicherheitskritische Bereiche.

Abstract

The European society increasingly relies on complex digital systems. In this context, we address the problem of working efficiently and reliably by using a large and complex set of interrelated technical documents of digital systems, where even human experts can hardly keep an overview or know all the details. This is particularly important in safety-critical domains such as electric vehicles, where such document sets are needed for specifying hardware-software interfaces correctly. Traditional methods of Information Retrieval do not allow for convenient access to the information in such documents, since

they require skills of using specific query languages, which, e.g., safety experts may not have.
Recent advances in Generative Artificial Intelligence (GenAI) and, more specifically, Large Language Models (LLMs) suggest that for the problem at hand, dialogues in natural language will be possible for accessing the content of such a document set. At the current state of the art, however, there is the major problem of “hallucinations” of LLMs, i.e., replying with pieces of information that are not based on factual knowledge but rather stem from more “creative” generations by the language model.
This is a major impediment to trust in such replies and, in general, such an approach, especially if it may be applied in a safety-critical domain such as electrical vehicles. Avoiding such hallucinations is key to reducing safety risks with such vehicles in the course of their development.
Trust is a complex psychological and sociological concept, and trustworthiness of a trustee is a major aspect of establishing trust. The significance of trust is not limited to the interpersonal domain; trust can also define the way people interact with technology. Trust in automation was transferred from trust relationships between humans. The truster is human, the trustee is an automated system.
Hence, we address ensuring trust through creating an LLM-based approach that is trustworthy. This is in line with the Regulation Artificial Intelligence Act (AI Act), which has just been approved by the European Parliament. The AI Act states: “The purpose of this Regulation is to promote the uptake of human centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law and the environment from harmful effects of artificial intelligence systems in the Union while supporting innovation and improving the functioning of the internal market. …”
Since the AI Act is a legal framework, our consortium includes legal experts, who will make sure that our research and development will really be in line with the AI Act and support its application. For making our approach human-centric, our consortium includes an expert on human-centered design that is diversity- and gender-sensitive.
Our proposed project includes research on a new approach to Requirements Engineering for specifying the needs in this regard. Our proposed research on Trust-related Requirements Engineering (RE) will be highly innovative by integrating quality attributes (in the sense of properties) of both systems (in our case, LLM-assisted systems for querying technical documentation) and data (in our case, primarily technical documentation) with each other, since properties of both system (such as transparency) and data (such as correctness) are relevant. At the current state of the art, only quality attributes of systems are taken into account in RE, but the importance of data and their qualities is ever increasing. Hence, for any non-trivial system based on data, quality attributes of both the system and its data are essential, and our new RE approach will take that into account in the development of trustworthy systems, in particular based on LLMs.
Based on concrete requirements on a specific LLM-assisted system to be defined according to this new approach, a specific tool owned by a scientific partner in this consortium (AERIALL, which supports free-to-use open-source models such as Mistral) will be further developed to satisfy these requirements for making it more trustworthy. Particular emphasis will be on avoiding hallucinations, while taking properties of both the system and its data into account.
This approach will be evaluated with real-world documents owned by the industrial partner in this consortium. In due course, it will be compared with the well-known ChatGPT approach from OpenAI, which may even be less performant than our proposed approach on custom information (in our case large and complex hardware-software interface documentation in a safety-critical domain), and especially not trustworthy.
With these innovations, the consortium will make a significant contribution to applying state-of-the-art AI technology to an important real-world problem, even taking safety-critical applications into account. Our overall approach is new and, therefore, not available yet, neither in Austria nor in Europe nor abroad. It will seamlessly link the AI Act with a new

approach to Requirements Engineering for trustworthy systems, and its use for an LLM-assisted approach to working with a large and complex set of technical documentation. This is particularly important for safety-critical domains.

Projektkoordinator

- Wirtschaftsuniversität Wien

Projektpartner

- Pro2Future GmbH
- Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur (IFZ)
- Robert Bosch Aktiengesellschaft