TrustBoost: Balancing flexibility and compliance in conversational AI systems

David Griol[1,2], Zoraida Callejas[1,2], Manuel Gil-Martín[3], Ksenia Kharitonova[1], Juan Manuel Montero Martínez[3], David Pérez-Fernández[1], Fernando Fernández-Martínez[3]

[1]Department of Software Engineering, University of Granada,
Periodista Daniel Saucedo Aranda S/N, 18071 Granada, Spain.
[2]Research Centre for Information and Communication Technologies, CITIC-UGR.
[3]Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group),
E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid,
Avda. Complutense 30, 28040 Madrid, Spain.

Correspondence: Zoraida Callejas zoraida@ugr.es

Abstract

Conversational AI (ConvAI) systems are gaining growing importance as an alternative for more natural interaction with digital services. In this context, Large Language Models (LLMs) have opened new possibilities for less restricted interaction and richer natural language understanding. However, despite their advanced capabilities, LLMs can pose accuracy and reliability problems, as they sometimes generate factually incorrect or contextually inappropriate content that does not fulfill the regulations or business rules of a specific application domain. In addition, they still do not possess the capability to adjust to users' needs and preferences, showing emotional awareness, while concurrently adhering to the regulations and limitations of their designated domain. In this paper, we present the TrustBoost project, which addresses the challenge of improving trustworthiness of ConvAI from two dimensions: cognition (adaptability, flexibility, compliance, and performance) and affectivity (familiarity, emotional dimension, and perception). The duration of the project is from September 2024 to December 2027.

1 Introduction

The term "Conversational AI" (ConvAI) has gained prominence in recent academic research, encompassing several NLP technologies like dialog systems, chatbots, and intelligent virtual assistants. These systems leverage artificial intelligence extensively to facilitate natural language conversations, offering users a seamless and intuitive way to access information, interact with services, and navigate data on the Internet, as well as their surrounding environment (Araujo and Bol, 2024; Casheekar et al., 2024; McTear, 2020; McTear et al., 2016)

In recent years, the development of LLMs has significantly transformed the landscape of conversational AI, offering unparalleled interaction flexibility. Unlike traditional rule-based or intent-based dialogue systems, LLMs can understand and generate human-like language that is less restricted because they are not as tied to specific training phrases. They have the ability to grasp context, adapt to diverse user inputs, and generate coherent responses in multiple conversational scenarios.

As for dialogue/interaction management, their pretraining on vast and diverse datasets enables them to generalize well to various domains, eliminating the need for explicit rule programming. Although rule-based systems often struggle to accommodate dynamic and evolving language patterns, LLMs contribute to a more natural and engaging user experience. On the opposite side of the spectrum, intent-based dialogue systems have the advantage of being completely compliant with business rules and domain restrictions due to their inherent structure and rule-based logic.

In intent-based systems, user interactions are categorized into predefined intents, each associated with a specific action or task. This structured approach provides a level of control and predictability that is advantageous to maintaining compliance. Regarding language understanding, intent-based systems are often specialized in specific domains or applications, enabling a focused understanding of user queries within a defined domain. In terms of dialogue management, this specialization contributes to a better alignment with domain-specific rules and restrictions, offering explainability.

Conversational LLMs are considered the future. However, they are still not widely adopted due to trustworthiness issues (Luna-Jiménez et al., 2022;

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 172–175

Kraus et al., 2021). Trust is crucial for user acceptance and engagement. Trust in AI systems is intricately linked to users' expectations of consistency, reliability, and adherence to established norms. Business rules serve as a set of guidelines that dictate the permissible behavior of the AI system within a given context. When a conversational system lacks compliance with domain restrictions, it introduces an element of unpredictability and inconsistency in its responses. This deviation from expected behavior can undermine users' trust.

Moreover, trust in ConvAI systems is intricately connected to interpretability: If users cannot understand the reasoning behind the system's decisions, it creates a perceived lack of transparency. From the perspective of the entity/company offering the conversational system to users, when a conversational system fails to align with business rules, the provider faces the risk of diminished user trust. This can lead to a decline in user engagement, increased user dissatisfaction, and potential reputational damage. In more extreme cases, they could even face legal and regulatory consequences.

2 Description

In TrustBoost, we assume that for ConvAI systems to be truly usable and reliable, they must adhere to the rules and restrictions of their designated domain as well as to adapt to their users' needs and preferences (cognition/performance branch of trustworthiness). With this aim, we will address the balance between flexibility and compliance by endowing ConvAI systems powered by LLMs the ability to comply with business/domain rules while simultaneously adapting to the needs of their users.

One of the primary challenges in incorporating rule-based constraints into the training and fine-tuning processes of LLMs is the discrepancy between the vast and diverse data sources typically used to train LLMs and the specific guidelines and regulations defined by organizations in the form of business rules. Training data may not adequately capture these intricacies, as LLMs lack the context of organizational policies and domain idiosyncrasies. Moreover, fine-tuning LLMs to adhere perfectly to business rules can be challenging due to the limited nature of fine-tuning datasets, which may not fully represent the complexity of these rules. Additionally, LLMs often generate outputs that are not easily interpretable, making it difficult to ensure that the generated content aligns with specific business rules and compliance requirements.

Due to the computational cost and resource constraints associated with training LLMs, their accessibility has remained limited to a select group of organizations with significant computational resources. To address this challenge, we propose to take advantage of the capabilities of 'big" pretrained LLMs to generate new resources, such as synthetic dialogues, that can be used to develop or fine-tune smaller and more efficient LLM models tailored to specific tasks and domains. These compact models would be able to run on average GPUs, expanding the reach of LLM technology to a wider range of organizations and applications. This approach would effectively break down the existing barrier to entry for LLMs, democratizing their use and fostering innovation across industries. Concurrently, we will devise automated evaluation procedures for the resources generated, minimizing manual review while ensuring the efficiency and reliability of the resources produced.

Within the framework of the TrustBoost project, we also aim to protect data governance and sovereignty by promoting the use of open source LLMs (to avoid sending data to third parties, unlike e.g. ChatGPT alternatives), together with techniques that make it possible to combine smaller models fine-tuned to specific tasks, ensuring the protection of personal data, lower hardware requirements for learning and deployment, and more efficient energy use.

In addition, in TrustBoost a truly flexible ConvAI system should dynamically tailor its responses based on the user's emotional cues (affective branch of trustworthiness). This adaptability involves adjusting the tone, language, content, and type of responses to align with the detected emotional state, thereby creating a more personalized and empathetic interaction. By doing so, the AI system can dynamically adapt the dialog flow to the identified emotional states, enhancing the overall user experience and building trust.

The challenge of recognizing the user's emotional state in ConvAI systems involves addressing multiple research goals. The first goal is the development of sophisticated natural language processing (NLP) techniques to extract and interpret emotional cues from various sources such as facial expressions, tone of voice, and word choice. Multimodal data integration, including spoken language, facial expressions, and body language, is crucial to enhance emotional recognition.

Recognizing the dynamic nature of emotions during interactions is also important. Contextualizing emotional cues within the conversation and the user's overall situation is vital for providing personalized and empathetic responses. However, current Conversational AI struggles to grasp the broader emotional landscape, facing limitations in understanding evolving user sentiments due to context window constraints. Overcoming these challenges requires advances in emotional state recognition and improving the contextual understanding capabilities of language models to enhance emotional intelligence in conversational interactions.

3 Main objectives

The main objective of TrustBoost is to find new methods for trustful Conversational AI balancing performance and affectiveness. In order to achieve this aim, we will address several research lines. First, enhancing user-awareness for trustworthy ConvAI. This research line aims to enhance user-awareness capabilities through the integration of techniques and algorithms from three perspectives: 1) integrating multimodal emotion recognition; 2) developing advanced NLP techniques to extract diverse emotional cues from user utterances; 3) focusing on memorability, enabling systems to identify and retain pivotal elements in dialogues.

Second, modeling trustworthiness in ConvAI. Our aim is to develop novel computational models that can predict and evaluate the trustworthiness of ConvAI systems based on performance, transparency and emotional intelligence.

Third, transitioning from intent-based dialogue systems to deep learning ConvAI. The most widely adopted technology for dialogue systems in industry is intent-based. We aim to transition from such technology to LLM-based ConvAI to achieve a more flexible interaction.

Fourth, generating conversational models compliant with business rules and domain restrictions. We will investigate how to generate the minimal model capable of adapting to well-defined business rules and domain restrictions, exploring ways of defining such rules, make them queryable through rule engines and coupling the engine with LLMs. In relation to this line, we plan to develop innovative approaches for generating new resources through prompt-based or instruction-based learning using LLMs, while simultaneously creating automated evaluation procedures for these resources.

Finally, TrustBoost advocates for the innovative exploration of strategies commonly used to detect hallucinations in conversational models, with the aim of evaluating whether individuals may encounter similar challenges. We will address this objective not only by using high-quality dialogue datasets avoiding biases, but also by assessing different techniques for hallucination mitigation: prompt engineering, self-refinement through feedback and reasoning, prompt tuning to adjust the instructions provided to a pre-trained LLM, decoding strategies, knowledge graphs, faithfulness based loss functions, or supervised fine-tuning.

4 Scientific and technical impact

TrustBoost foreseen advances affect the architecture of ConvAI systems by proposing the integration of LLMs and additional components to improve the interaction context, compliance with associated business rules, multimodal interaction, and user adaptation. The project will provide scientific impact related but not limited to: transitioning from intent-based dialogue systems to LLM-based conversational AI provided added flexibility while maintaining compliance; generating quality open-access dialogue resources for the training of these models in Spanish and English for multiple domains; new tools and platforms to develop and evaluate ConvAI systems; developing new methods for the integration of smaller language models that meet the requirements associated to data protection, provide accuracy results comparable to larger models, and allow reducing hardware and energy requirements for their deployment; generating new methods for mitigating hallucinations and explainability, fostering trust; reducing the number of responses factually incorrect or contextually inappropriate that do not fulfill the regulations or business rules of a specific application domain; new methods for understanding and responding to user emotional cues; new techniques for integrating the emotional cues into LLM-based conversational AI; and novel and more trustworthy ConvAI models that are user-aware, and emotionally interactive.

By enhancing user experiences, building trust, and promoting explainable and adaptable interactions with ConvAI systems, TrustBoost can contribute to a more positive and supportive technological environment, ultimately benefiting the well-being of the population.

Acknowledgments

The TrustBoost project has received funding from MICIU/AEI/10.13039/501100011033 and from FEDER, UE. It is a coordinated project by a multidisciplinary team from the Universidad Politécnica de Madrid (UPM) and University of Granada (UGR), with two subprojects that address TrustBoost's objectives: "Enhancing Trustworthiness in Conversational AI through Multimodal Affective Awareness" (TrustBoost-UPM, ref. PID2023-150584OB-C21), and "Breaking the duality of conversational AI: going beyond guided conversations while ensuring compliance with domain rules and constraints" (TrustBoost-UGR, ref. PID2023-150584OB-C22).

References

T. Araujo and N. Bol. 2024. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1):100030.

A. Casheekar, A. Lahiri, K. Rath, K. Sanjay Prabhakar, and K. Srinivasan. 2024. A contemporary review on chatbots, ai-powered virtual conversational agents, chatgpt: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632.

M. Kraus, N. Wagner, Z. Callejas, and W. Minker. 2021. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836.

C. Luna-Jiménez, S.L. Lutfi, and F. Fernández-Martínez. 2022. Measuring trust at zero-acquaintance: A cross-cultural study between malaysians and hungarians. In *Proc. of 26th International Conference on Intelligent Engineering Systems (INES)*, page 000267–000272.

M.F. McTear. 2020. *Conversational AI. Dialogue systems, Conversational Agents, and Chatbots*. Morgan and Claypool Publishers.

M.F. McTear, Z. Callejas, and D. Griol. 2016. *The Conversational Interface: Talking to Smart Devices*. Springer.