DOI: 10.11576/dataninja-1160

Linguistic-Based Reflection on Trust Calibration in Conversations with LLM-Based Chatbots

Milena Belosevic MILENA.BELOSEVIC@UNI-BIELEFELD.DE
German Linguistics, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Hendrik Buschmeier HBUSCHME@UNI-BIELEFELD.DE
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Abstract

This paper presents a linguistic approach to trust in human conversations with LLM-based chatbots. Using the concept of trust calibration [1] as a starting point, we aim to address the question of how to increase user AI literacy and prevent misuse of as well as overtrust in the information provided by LLM-based chatbots in educational contexts. We propose a linguistic-based model of trust calibration that supports users in adopting a critical perspective on trust calibration and controlling their trust level. The method combines previous studies on trust in human interaction, specifically linguistic trust cues displayed by human trustors to indicate their level of trustworthiness in naturally occurring contexts [see 2] with studies on proactive human-computer interaction [3] and the social influence of conversational agent's embodiment in educational contexts [4].

Keywords: trust calibration; linguistic trust cues; LLM-based chatbots

1. Background

Trusting information provided by large language models (LLMs) has received growing attention with the advent of LLM-based chatbots. Currently, users have high expectations regarding the desired capabilities of LLM-based chatbots. As Wang et al. [5] note, "users expect LLMs to be multifaceted, capable of accurately solving complex professional tasks, and rich in providing personalized or novel responses". Such expectations can have serious consequences, especially in educational contexts where students typically lack literacy in artificial intelligence (AI). In this paper, we propose and test a linguistic method for trust calibration in educational settings, which supports students in reflecting on and controlling their trust in a chatbot's responses, thus increasing their AI literacy, and agency [6]. Previous studies have focused on adjusting the level of trust in LLMs by training models to display confidence levels [7]. Another group of studies is concerned with developing specific prompting strategies [8], or training the models to elicit the appropriate level of trustworthiness [9] or conducting user studies to elicit users' experience and expectations regarding the desired design of LLM-based chatbots [10]. However, although "the choice of words is a vehicle for establishing trust in interpersonal online communication – regardless of whether it is written or spoken and whether the interaction is with another human or an artificial interlocutor" [11], to our knowledge, trust calibration has not been approached from linguistic perspectives [but see 12].

Starting from the hypothesis that linguistic trust cues that trustees use to indicate their trustworthiness in human interaction (e.g., markers of (un)certainty, referring to experts and numbers, lexical alignment, etc.) can be accidentally generated by LLMs as next words in particular contexts, our model complements previous research on trust in LLMs by focusing on how these cues can help students to adopt a critical stance towards the information provided by LLM-based chatbots and support them to engage in critical reflection on how to control their trust in LLM-based chatbots. To this end, we introduce a new phase of trust calibration, the 'reflection phase', which complements the existing aspects of trust calibration (e.g., overtrust, undertrust, etc. [13]) by including conversational agents in students' interaction with LLM-based chatbots.

2. Methodology

Our model is designed for the context of higher education [14, 15] and is based on the following idea: A conversational agent designed to act as a learning assistant helps students remain aware that LLMs are merely next-word predictors that should not generally be trusted in the same manner as we trust humans. By providing optional assistance in the form of (non)verbal dialogue actions, the conversational agent supports students' critical reflection and control over their trust in the chatbot's responses [4]. Suppose, for example, that the chatbot's response contains the verb 'understand', such as in 'I *understand* what you are trying to say.' This should be regarded as problematic because this linguistic unit can be associated with a set of trust cues denoting the orientation toward the trustor's needs and goals in human interaction [16, 17].

We argue that such linguistic units should not be perceived as trust cues but should instead serve to motivate and support users' critical reflection on the reliability of the response. This can be achieved by designing conversational agents to display verbal, nonverbal, and prosodic-acoustic metacommunicative expressions (e.g., distance markers/quotation marks) embedded in the agent's proactive dialogue actions [3], such as notifications, suggestions, or interventions. Using specific conversational acts (e.g., acknowledgment, see [18]), conversational agents can assist students in assessing the trustworthiness of the chatbot's response based on the verbal trustworthiness cues displayed in the response. Importantly, the agent first offers help, but students can decide whether they want the agent's assistance.

3. Case study

To design a conversational agent capable of providing proactive dialogue actions that help students control their trust in the chatbot's responses, it is first necessary to identify which linguistic units are perceived as linguistic trust cues in human interactions with LLM-based chatbots. To this end, we conducted a rating study to test the influence of one type of linguistic units that potentially influence users' trust, namely grounding acts [18]. As noted by Chiesurin et al. [19], current LLM-based dialogue systems usually guess what the user intended instead of leveraging grounding acts, which may lead to miscalibrated trust and overconfidence. We tested the following two hypotheses: (H1) The responses in the 'baseline' condition will receive lower trustworthiness ratings than in two alternative conditions (see also [20]). (H2) The perceived trustworthiness is higher in the 'anthropomorphic' than in the 'grounding act' condition (contrary to [21]).

In a within-subject design study, students ($N = 32$; 17 female, 15 male; 14 German native speakers, 12 bilingual, 6 non-native speakers of German; age: $\mathrm{M} = 25.96, \mathrm{SD} = 4, \mathrm{Mdn} = 26$) were exposed to items from these three conditions[1]. The acknowledgment and the anthropomorphic condition comprised the *Other-Acknowledgment* speech-act pattern, specifically the *inform* → *ackn+mrequest* pattern [22], where a student (other) presents a math task (inform) selected from the MathDial dataset [23] and a virtual tutor (ChatGPT 3.5) was prompted to respond by acknowledging the information (*ackn*). In the acknowledgment condition, the acknowledgment is verbalized by 'In Ordnung', 'Alles klar', etc. and followed by a request for clarification of some part of the information to verify understanding (*mrequest*). In the anthropomorphic condition, the math task presented by the student was followed by the tutor's acknowledgment and a follow-up question verbalized by anthropomorphic verbs (e.g., 'understand'). The baseline condition comprised the tutor's direct response to the student's task. The perceived trustworthiness of the tutor's response was measured indirectly by having participants rate the following statements: "The virtual chatbot tutor can help the student." The participants responded by selecting a value on a four-point Likert scale. Both hypotheses were confirmed by statistical analyses: The responses in the baseline condition received lower mean trustworthiness ratings ($\mathrm{M} = 1.93$) than in the anthropomorphic ($\mathrm{M} = 3.23$) and the acknowledgment condition ($\mathrm{M} = 2.98$). As indicated by these values, the perceived trustworthiness received higher ratings in the anthropomorphic than in the acknowledgment condition (see H2). The differences between the three conditions are statistically significant (Friedman Test, $\chi = 28.79, p < 0.001$).

4. Conclusions and future work

The results obtained in the questionnaire study can be used as a starting point for providing recommendations for designing communication strategies for trustworthy AI [8, 10]. In the next step, we will use these results to test whether enriching grounding acts with nonverbal aspects affects students' perceived trust in the chatbot's responses. To this end, we plan to include a social robot in the conversation with LLM-based chatbots.

[1] See the supplementary material for study design, dataset, and results: https://doi.org/10.17605/OSF.IO/FYQ3P.

References

[1] Bonnie M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27:527–539, 1987. doi: 10.1016/S0020-7373(87)80013-5.

[2] Mie Femø Nielsen and Ann Merrit Rikke Nielsen. *Revisiting Trustworthiness in Social Interaction*. Routledge, New York, NY, USA, 2022. doi: 10.4324/9781003280903.

[3] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836, 2021. doi: 10.1109/ACCESS.2021.3103893.

[4] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, pages 1882–1887, Sapporo, Japan, 2012.

[5] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. Understanding user experience in large language model interactions. *arXiv:2401.08329*, 2024. doi: 10.48550/arXiv.2401.08329.

[6] Don Passey, Miri Shonfeld, Lon Appleby, Miriam Judge, Toshinori Saito, and Anneke Smits. Digital agency: Empowering equity in and through education. *Technology, Knowledge and Learning*, 23:425–439, 2018. doi: 10.1007/s10758-018-9384-x.

[7] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.48550/arXiv.2012.14983.

[8] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. "as an ai language model, I cannot": Investigating LLM denials of user requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Honululu, HI, USA, 2024. doi: 10.1145/3613904.3642135.

[9] Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. Verifiable by design: Aligning language models to quote from pre-training data. *arXiv:2404.03862*, 2024. doi: 10.48550/arXiv.2404.03862.

[10] Mateusz Dubiel, Sylvain Daronnat, and Luis A Leiva. Conversational agents trust calibration: A user-centred perspective to design. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6, Glasgow, UK, 2022. doi: 10.1145/3543829.3544518.

[11] Regina Jucks, Gesa A. Linnemann, Franziska M. Thon, and Maria Zimmermann. Trust the words: Insights into the role of language in trust building in a digitalized world. In Bernd Blöbaum, editor, *Trust and Communication in a Digitized World: Models and Concepts of Trust Research*, pages 225–237. Springer, Cham, Switzerland, 2016. doi: 10.1007/978-3-319-28059-2_13.

[12] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil, 2024. doi: 10.48550/arXiv.2405.00623.

[13] Ewart J. De Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12:459–478, 2020. doi: 10.1007/s12369-019-00596-x.

[14] Melissa Donnermann, Philipp Schaper, and Birgit Lugrin. Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. *Frontiers in Robotics and AI*, 9:831633, 2022. doi: 10.3389/frobt.2022.831633.

[15] Thomas Beelen, Khiet Truong, Roeland Ordelman, Ella Velner, Vanessa Evers, and Theo Huibers. A child-friendly approach to spoken conversational search. In *Proceedings of the 2nd Workshop on Mixed-Initiative Conversational Systems (MICROS)*, Atlanta, GA, USA, 2022.

[16] Pavla Schäfer. *Linguistische Vertrauensforschung: Eine Einführung.* de Gruyter, Berlin, Germany, 2016. doi: 10.1515/9783110451863.

[17] Martha Kuhnhenn. *Glaubwürdigkeit in der politischen Kommunikation Gesprächsstile und ihre Rezeption.* UVK, Konstanz, Germany, 2014.

[18] David R. Traum and Elizabeth A. Hinkelmann. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8:575–599, 1992. doi: 10.1111/j.1467-8640.1992.tb00380.x.

[19] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada, 2023. doi: 10.18653/v1/2023.findings-acl.60.

[20] Gesa Alena Linnemann and Regina Jucks. ‘Can I trust the spoken dialogue system because it uses the same words as I do?’ – Influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers*, 30:173–186, 2018. doi: 10.1093/iwc/iwy005.

[21] Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M. Bender. From “AI” to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil, 2024. doi: 10.48550/arXiv.2404.16047.

[22] David G. Novick and Stephen Sutton. An empirical model of acknowledgement for spoken-language systems. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 96–101, Las Cruces, NM, USA, 1994. doi: 10.3115/981732.981746.

[23] Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, 2023. doi: 10.18653/v1/2023.findings-emnlp.372.