Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

Mapping and Calibrating User Trust with LLMs: First Steps Towards Developing a Framework for Shaping Trust

Samuel Hill[1] [0009-0004-9621-0033], Joy Belgassem[2] (0009-0005-5962-2554), Felix Nadolni[2]

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Trippstadter Strasse 122, 67663 Kaiserslautern, Germany
samuel.hill@dfki.de
[2] Bundesdruckerei GmbH, Kommandantenstr. 18, 10969 Berlin, Germany
nataliejoy.belgassem@bdr.de
felix.nadolni@bdr.de

Abstract. Trust in Large Language Models (LLMs) is critical for ethical and effective deployment, especially in high-stakes public sector contexts. This study combines a literature review and a qualitative user study with public administration professionals to explore how user trust in LLMs can be mapped and calibrated. The result is a Trust Areas/Trust Dimensions (TA/TD) framework that identifies key factors influencing trust, including accuracy, transparency, privacy, and ethical considerations. The framework supports trust calibration by aligning user expectations with system capabilities and informs future design and governance strategies. It offers a structured, adaptable tool for evaluating and guiding trust in LLMs across evolving technological and societal landscapes.

Keywords: Trust in AI, Trustworthy AI, Large Language Models

1 Introduction and Methods to Framing Trust

In democratic societies, trust in public institutions is foundational. Governments are not only tasked with regulating emerging technologies like artificial intelligence (AI) but are increasingly adopting them to enhance public services, decision-making, and regulatory enforcement. This dual role places governments in a uniquely high-accountability position, making them an ideal and extreme use-case for studying trust in AI systems, particularly large language models (LLMs).

The deployment of AI in public administration directly impacts citizens' rights, access to services, and perceptions of fairness, privacy, and justice. Misuse or misalignment of AI tools in this context risks eroding public confidence in democratic institutions. Therefore, mapping and calibrating trust in government use of AI is a critical challenge, both technically and ethically.

Trust in AI is multifaceted and context-dependent. It involves balancing overtrust, which can lead to blind reliance on opaque systems, and undertrust, which may hinder

Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium
on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

beneficial adoption. This balance is captured in the concept of calibrated trust. ensuring that users' confidence in AI systems aligns with the systems' actual capabilities and limitations. In public administration, where decisions can have profound societal consequences, achieving calibrated trust is especially critical.

LLMs present unique challenges due to their probabilistic nature and lack of transparency. Their outputs can vary unpredictably, making it difficult for users to assess reliability. This unpredictability underscores the need for structured frameworks that help users understand, evaluate, and recalibrate their trust in these systems over time.

Our research focuses on this high-stakes context of public administration to explore how trust is formed, maintained, and adjusted in human–LLM interactions. Through literature reviews, interviews with public sector professionals, and thematic analysis, our aim led to defining a Trust Framework of Trust Areas (TAs) and Trust Dimensions (TDs) that can guide future design interventions and governance strategies to map and calibrate user trust when interacting with these highly complex and necessary systems.

2 Literature Review, Evolving Perspectives on Trust

At the outset of this project, our literature review was designed to establish a foundational understanding of trust by tracing its conceptual evolution across disciplines, from social psychology and philosophy to automation and information systems. This historical trajectory provided a robust baseline for developing a Trust Framework applicable to emerging technologies like LLMs. While many of the trust concepts we explored are rooted in earlier scholarship, they remain relevant and adaptable as the technological landscape evolves. We acknowledge that more recent work on trust in AI and LLMs has emerged since our initial review, and future iterations of the Trust Framework should integrate these developments to ensure continued relevance.

The review synthesized insights from over 70 key publications and organized them into thematic chapters. Early definitions of trust emphasized vulnerability and expectation (e.g., Mayer et al., 1995; Rousseau et al., 1998), while later models introduced dimensions such as performance, process, and purpose (Söllner et al., 2012), and humanness in system design (Lankton et al., 2015). Trust in automation and information systems was shown to hinge on calibration, aligning user expectations with system capabilities, and on the perceived reliability and transparency of outputs.

As the review progressed toward AI and LLMs, it highlighted unique challenges. LLMs operate as probabilistic black boxes, often producing outputs without clear rationale. This opacity can lead to overtrust or undertrust, especially when users rely on past performance or lack awareness of system limitations. Studies on trust calibration (e.g., Zhang et al., 2020; Benz & Rodriguez, 2024) suggest that conveying confidence levels or selectively withholding uncertain outputs can help users make informed decisions. Recent literature also explores language as a trust mediator, with linguistic politeness and clarity shown to foster trust in human–machine interactions. Moreover, empirical studies on LLMs reveal mixed results: while some users perceive AI-generated content as high quality (Zhang & Gosline, 2023), others remain skeptical due to hallucinations and lack of source attribution (Döbler et al., 2024). Importantly, the

Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium
on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

review underscores that trust in LLMs is not monolithic, it is shaped by context, user experience, perceived risk, and the design of the system itself.

3 Qualitative User Study on Trust in LLMs

To complement the literature review with real-world exploratory perspectives, we conducted semi-structured interviews with 21 professionals across various public sector roles, including educators, IT specialists, engineers, and government officials.
LLMs are primarily used for text-related tasks such as summarization, drafting emails, generating reports, and translating documents. Many participants described LLMs as "assistants" that help reduce workload and improve efficiency, especially in administrative and knowledge-heavy environments. For example, one volunteer services advisor used LLMs to shorten lengthy reports and format event programs, while a digitalization strategy chief emphasized their utility in extracting key points from convoluted legal texts.

In educational settings, teachers used LLMs to support students in understanding complex texts and for translating materials while simultaneously concerned about overreliance and the erosion of critical thinking skills. Across interviews, a common theme was the human-in-the-loop approach, where LLM outputs are reviewed, edited, or used as a starting point rather than final products.

Trust in LLMs was shaped by several interrelated dimensions. *Accuracy and Reliability:* Users consistently emphasized the need for verifiable outputs and factual correctness through source citation and utilizing retrieval-augmented generation (RAG) systems. In statistical offices, for instance, LLMs were configured to point users to existing tables rather than generate new content, minimizing hallucination risks. *Transparency and Explainability:* Interviewees expressed a desire for LLMs to be more transparent about how outputs are generated. This was especially important in legal and compliance contexts, where decisions must be traceable and justifiable. *Data Privacy and Security:* Concerns about sensitive data were widespread. Several participants preferred self-hosted or open-source models (e.g., Llama, Mistral) over commercial cloud-based solutions. Internal guidelines and secrecy procedures were often cited as necessary safeguards. *Ethical and Social Considerations:* Participants highlighted the importance of digital sovereignty, bias mitigation, and ethical education. Some worried about the ecological impact of LLMs or the potential for job displacement, while others stressed the need to maintain human judgment in decision-making processes.

Adoption of LLMs in the public sector is uneven and often hindered by infrastructure limitations, legal uncertainties, and organizational resistance. Teachers, for example, cited poor Wi-Fi and lack of digital tools as barriers to effective integration. Others noted generational differences in openness to AI, with younger employees more willing to experiment. Conversely, training and education emerged as key enablers. Many interviewees were involved in developing internal guidelines, conducting workshops, or creating e-learning modules to prepare colleagues for AI integration. The importance of clear usage policies, ethical literacy, and supportive onboarding was repeatedly emphasized.

Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium
on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

Interviewees identified a range of potential future applications for leveraging LLMs not just for efficiency, but also for accessibility, personalization, and knowledge democratization, including: 1) Automated job description and applicant review, 2) Legal document summarization and decision support, 3) Internal knowledge navigation (e.g., compliance databases), 4) Personalized tutoring and differentiated learning materials, and 5) AI-powered chatbots for citizen services and statistical queries.

4 Trust Framework

The Trust Framework and the corresponding TAs and TDs were developed to analyze user trust in LLM systems, drawing from both the literature review and the qualitative user study. It offers a structured lens for understanding how trust is built, sustained, or misaligned. Designed to be dynamic, the framework remains responsive to ongoing shifts in trust-related scholarship and the fast-evolving capabilities and contexts of LLM technologies.

To develop the Trust Framework, affinity mapping and clustering methods were used to identify patterns and overlapping concepts between theoretical trust concepts and user insights on LLM applications, which were distilled into distinct, independent TDs. These were grouped into five overarching TAs: *USER, MODEL, INTERFACE, DATA,* and *ETHICS,* with each TA encompassing specific TDs that describe either inherent qualities (e.g., AI Literacy, Performance) or relational dynamics (e.g., Level of Automation, Explainability). Several TDs span multiple TAs, reflecting the layered and interconnected nature of trust in LLM interactions.

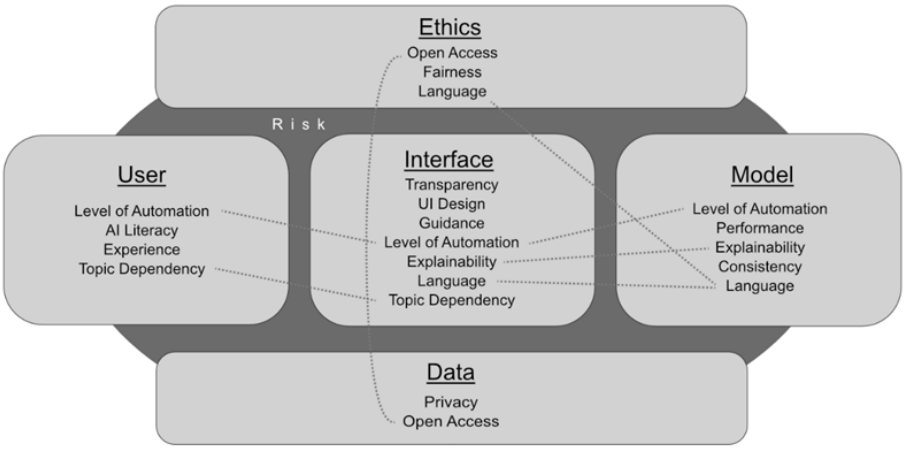

Fig. 1. A diagram showing Trust Areas (TAs) and Trust Domains (TDs). Dashed lines connect instances where specific TDs appear in multiple TAs, highlighting the framework's interconnected and layered structure.

Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

TDs such as *Explainability, Consistency,* and *Transparency* directly address the black-box characteristics of LLMs and their unpredictable outputs. These dimensions respond to concerns raised in both the literature and interviews about the opacity of model reasoning and the variability of outputs.

USER-related TDs, including *AI Literacy, Experience,* and *Topic Dependency,* capture how users form and adjust trust based on their interaction history, domain expertise, and understanding of LLM capabilities. These TDs emerged strongly from interview data, where users described trust as contingent on their familiarity with both the task and the system.

INTERFACE-related TDs such as *UI Design* and *Guidance* reflect how LLMs communicate their capabilities and limitations. These TDs are critical for shaping user expectations and emerged as key leverage points for trust calibration in both theoretical models and user feedback.

The DATA and ETHICS TAs incorporate TDs like *Privacy, Open Access,* and *Fairness,* which are central to public sector deployment and regulatory compliance. These TDs align with broader concerns in the literature about responsible AI and were frequently cited by interviewees as prerequisites for trust.

Compared to existing models, the TA/TD Framework offers several distinct advantages, for example: 1) It integrates multi-actor perspectives. While developed from end-user insights, the framework implicitly maps roles for providers, regulators, and developers, enabling actor-specific trust strategies and facilitating cross-stakeholder dialogue. 2) It supports trust calibration. By identifying dimensions that contribute to overtrust or undertrust, the framework provides a vocabulary and structure for designing targeted interventions, whether through interface design, user education, or policy mechanisms. 3) It enables comparative analysis. The framework allows for cross-model and cross-interface evaluations, supporting empirical studies and design decisions that account for variation in LLM behavior and user expectations. 4) It aligns with policy needs. The framework is well-suited for integration into regulatory frameworks such as the EU AI Act, offering actionable dimensions for compliance, governance, and public accountability.

5 Discussion and Future Research

Although our presented TA/TD Framework reflects a specific point in time, it is designed to be adaptable to ongoing technological and societal change. In its current form, it provides a baseline for evaluating how TAs/TDs interrelate, and how these interdependencies influence the mapping and calibration of trust. Future research will pursue questions that emerge from this foundation, including how trust can be measured, balanced, or redistributed across dimensions. This includes co-design interventions to address these questions practically.

Moreover, the framework encourages deeper integration between literature synthesis and qualitative insights, ensuring that both theoretical and experiential understandings of trust evolve together. The TA/TD Framework is a starting point, not a conclusion, calling for ongoing refinement, interdisciplinary collaboration, and responsiveness to

emerging regulations like the EU AI Act. Ultimately, this work aims to ensure that trust in LLMs is not only understood but also actively shaped and sustained.

Acknowledgments. The authors would like to acknowledge the support of this project by Bundesdruckerei GmbH and Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI).

Disclosure of Interests. The authors are employed by Bundesdruckerei GmbH and by Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), which may benefit from the results presented in this paper.

Appendix

Tables 1-thru-5 present the short-form definitions of each TA/TD within the Trust Framework which were developed alongside the more detailed descriptions.

Table 1. Trust Area: USER.

USER: This cluster focuses on the user's interaction, perception, and relationship with the system. It emphasizes both the practical and emotional aspects of trust, as well as psychological and behavioral dimensions.	
Level of Automation	Distribution of cognitive work in human-machine-collaboration.
AI Literacy	A user's theoretical understanding of LLMs and resulting calibrated expectation of the results from an interaction.
Experience	A user's practical knowledge of LLM systems gained through interaction.
Topic Dependency	The user's level of expertise on the subject of interaction.

Table 2. Trust Area: MODEL.

MODEL: This cluster concerns the technical capabilities and limitations of the LLM itself, as products of hardware, model architecture and training methods. The organization facilitating and/or hosting the model is also part of this cluster.	
Level of Automation	[see TA User]
Performance	A system's capabilities to provide reliable, accurate, and useful outputs.
Explainability	The communication of a system's reasoning processes, decision-making logic, and the factors influencing its outputs, including potential limitations and sources of uncertainty.
Consistency	A systems capability to produce output of stable quality.
Language	A system's capacity to produce output in varying linguistic forms and the ability to choose an appropriate mode for a given interaction.

Table 3. Trust Area: INTERFACE.

Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

INTERFACE: This cluster deals with how users interact with the system, emphasizing the design principles that influence trust, prompting, and partly fine-tuning. This also includes how a system communicates its capabilities.	
Transparency	Communication of a system's capabilities and fallibilities, known biases, and uncertainties.
UI Design	The virtual locale and mode of communication between user and model.
Guidance	A measure of determination of user-model interaction, i.e. how strictly it follows a predefined structure.
Level of Automation	[see TA User]
Explainability	[see TA Model]
Language	[see TA Model]
Topic Dependency	[see TA User], with the addition of *aligning* the model to the user's knowledge.

Table 4. Trust Area: DATA.

DATA: This cluster focuses on the data used to train and operate the LLM, including questions around copyright, as well as the handling of user inputs with respect to privacy and data security questions.	
Privacy	Secure and responsible handling of data.
Open Access	A significant factor regarding the "black box" characteristics of a model; parallel to the discussion around Open Source / Open Data.

Table 5. Trust Area: ETHICS.

ETHICS: This cluster addresses the moral and societal implications of LLMs.	
Open Access	[see TA Data]
Fairness	The absence of discrimination and biases in LLM behavior.
Language	[see TA Model]

References

1. Benz, N., Rodriquez, M.: Human-aligned calibration for AI assisted decision making. arXiv preprint arXiv:2306.00074 (2024). https://arxiv.org/abs/2306.00074
2. Döbler, M., Mahendravarman, R., Moskvina, A., Saef, N.: Can I trust you? LLMs as conversational agents. In: Deshpande, A., Hwang, E., Murahari, V., Park, J.S., Yang, D., Sabharwal, A., Narasimhan, K., Kalyan, A. (eds.) Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pp. 71–75. Association for Computational Linguistics, St. Julians (2024).
3. Lankton, N., McKnight, D., Tripp, J.: Technology, humanness, and trust: rethinking trust in technology. J. Assoc. Inf. Syst. 16 (10), 880–918 (2015). https://doi.org/10.17705/1jais.00411
4. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. 20(3), 709–734 (1995). https://www.jstor.org/stable/258792

5. Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C.: Not so different after all: a cross-discipline view of trust. Acad. Manag. Rev. 23 (3), 393–404 (1998). https://doi.org/10.5465/amr.1998.926617
6. Söllner, M., Hoffmann, A., Hoffmann, H., Wacker, A., Leimeister, J.M.: Understanding the formation of trust in IT artifacts. In: Proceedings of the 33rd International Conference on Information Systems (ICIS 2012). Association for Information Systems, Orlando (2012). https://aisel.aisnet.org/icis2012/proceedings/HumanBehavior/11
7. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. ACM, New York (2020). https://doi.org/10.1145/3351095.3372852
8. Zhang, Y., Gosline, R.: Human favoritism, not AI aversion: people's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. Judgm. Decis. Mak. 18 (41) (2023). https://doi.org/10.1017/jdm.2023.37