WhiteBox

Symposium on Explainability

LOEWE Exzellente Forschung für Hessens Zukunft · Centre for Cognitive Science · TECHNISCHE UNIVERSITÄT DARMSTADT

Sven Schultze

Trust Calibration in Large Language Models with XAI

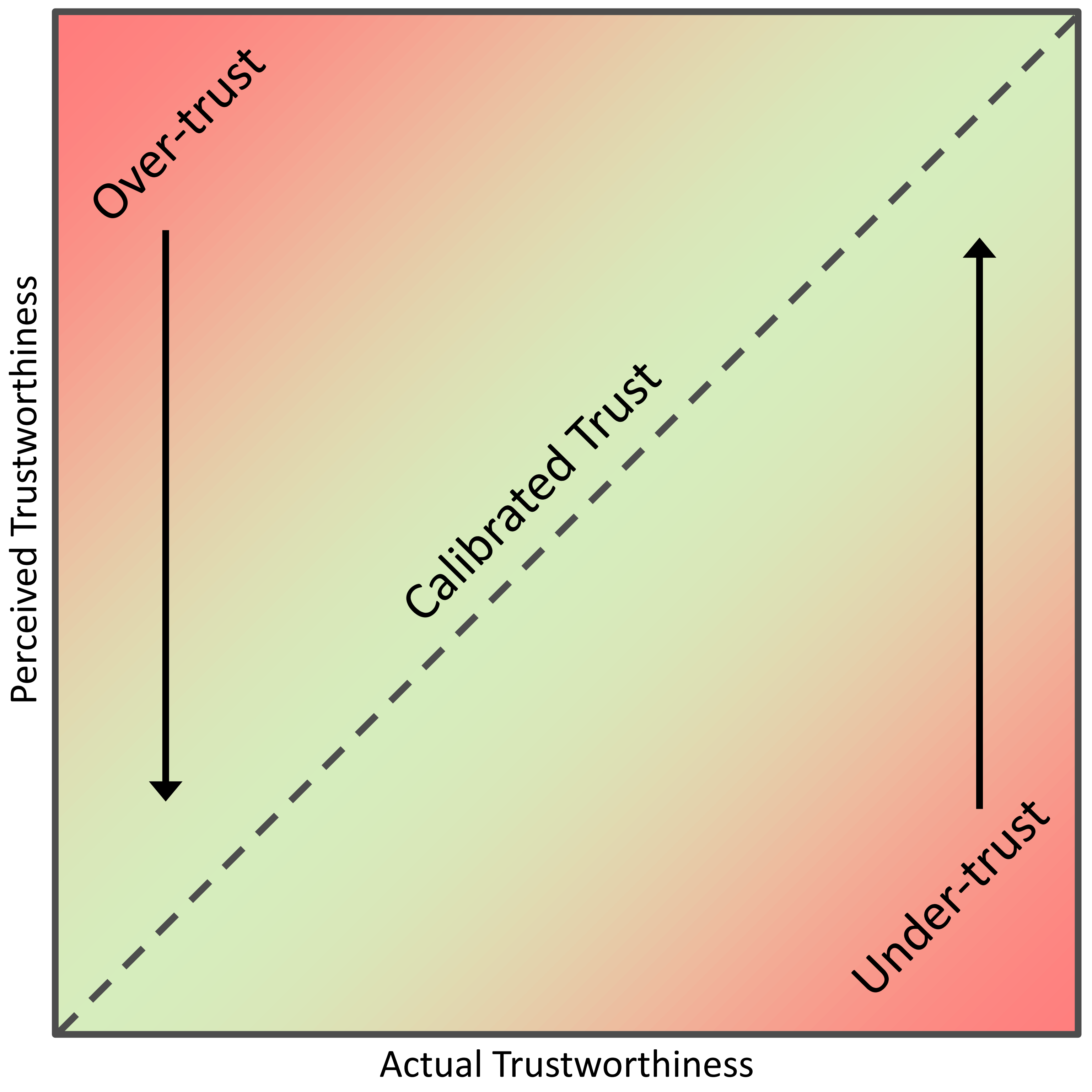
Large Language Models (LLMs) have been observed to exhibit a phenomenon known as hallucination, where they generate seemingly plausible but factually inaccurate responses when presented with information that is outside the scope of their training data.[1] Ascertaining the reasons for such responses can be challenging, making it arduous for users to determine the reliability of the model.

Explainable AI (XAI) can address this issue by providing users with insights into the model's decision-making process. XAI techniques can explain how the model arrived at its response, identify areas prone to hallucination, and increase transparency in decision-making. This can lead to improved accuracy, accountability, and user trust in AI technology.

Research Questions

1) How can we develop reliable methods for detecting when LLMs are prone to hallucination, and what corrective measures can be taken in response?
2) What are the most effective XAI techniques for providing users with insights into the decision-making process of LLMs?
3) Can XAI techniques influence users of LLMs, to reduce overtrust when the LLM is wrong and reduce undertrust when it is right?

Trust Calibration



When users overestimate the capabilities of an AI system, they may rely too heavily on its recommendations or decisions, leading to potential errors or biases. Conversely, if users underestimate the system's capabilities, they may not utilize it effectively, leading to missed opportunities.[2]

Methods

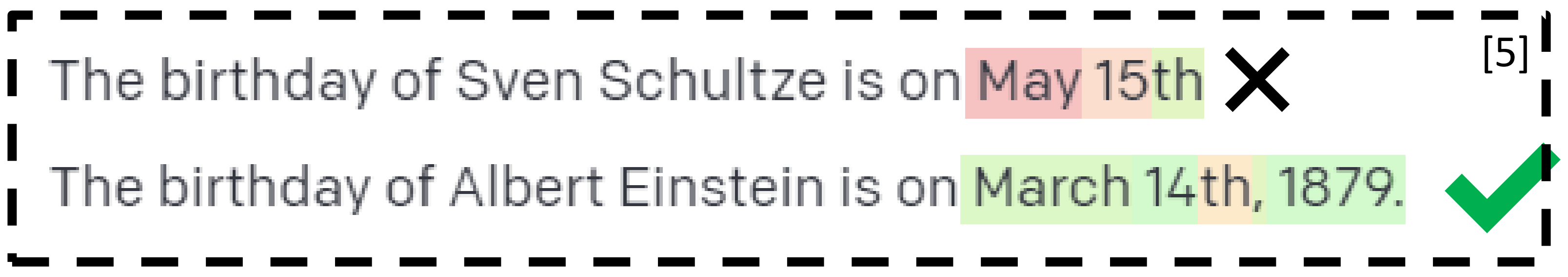
1) Out-of-Distribution Detection[3]

This involves detecting when the model is presented with data that is outside of its training distribution. By analyzing the model's response to this type of data, it is possible to identify areas where the model may generate inaccurate responses or hallucinations.

2) Counterfactual Explanations[4]

This involves generating alternative scenarios that could have led to a different decision by the model. By comparing the actual decision made by the model to the counterfactual explanations, it is possible to identify areas where the model may have generated inaccurate responses or hallucinations.

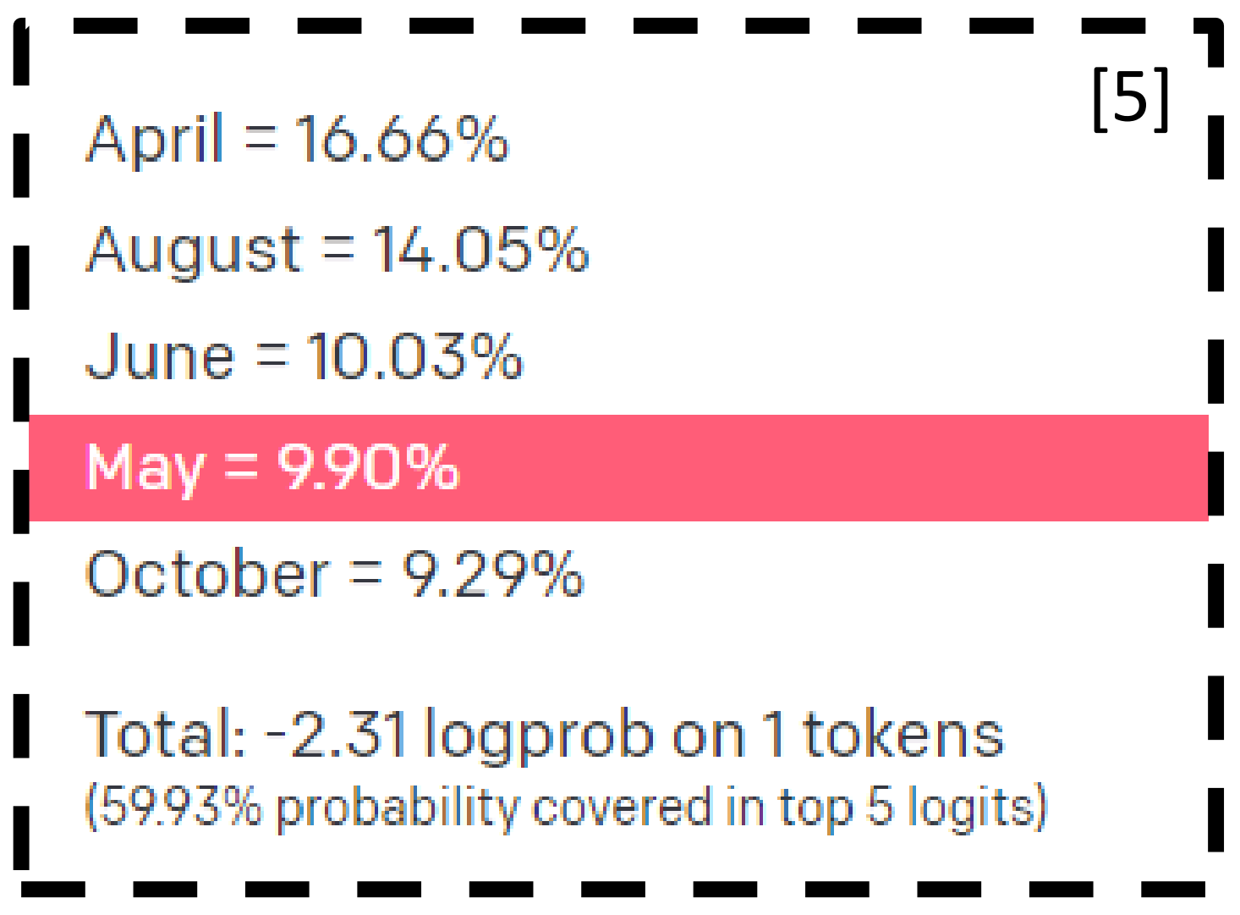
3) Model Confidence



The birthday of Sven Schultze is on May 15th ✗

The birthday of Albert Einstein is on March 14th, 1879. ✓

[5]



[5]

April = 16.66%
August = 14.05%
June = 10.03%
May = 9.90%
October = 9.29%

Total: -2.31 logprob on 1 tokens
(59.93% probability covered in top 5 logits)

This involves analyzing the confidence scores generated by the AI model for each response it generates. If the model is generating responses with low confidence scores, and there are alternative completions of the same category (like other Months) it may be an indication that the model output is not trustworthy.

Next Steps

1. Apply Methods to OpenChatKit Model
2. Design UI Prototype to combine Explanations
3. Improve UI in iterative Human-Centered Design Process
4. Evaluate influence of UI on user trust calibration

[1] Ji, Ziwei, et al. "Survey of hallucination in natural language generation." ACM Computing Surveys (2022).
[2] De Visser, E. J., et al. *Towards a theory of longitudinal trust calibration in human–robot teams.* International journal of social robotics 12.2 (2020)
[3] Ren, J., et al. *Out-of-Distribution Detection and Selective Generation for Conditional Language Models.* NeurIPS 2022 Workshop on Robustness in Sequence Modeling.
[4] Dandl, S. and Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2nd ed., 2022, https://christophm.github.io/interpretable-ml-book.
[5] Images taken from OpenAI GPT-3 Playground: https://platform.openai.com/playground/p/

TU Darmstadt / 27.03.2023
LOEWE-Schwerpunkt WhiteBox