Chair of Explainable AI-Based Business Information Systems
Prof. Dr. Ulrich Gnewuch

Bachelor's/Master’s Thesis: Uncertainty-Triggered Intelligent Explanations in LLM-based Data Assistants

Supervisor: Ana-Maria Sîrbu (anamaria.sirbu@uni-passau.de)
Start date: as soon as possible

Motivation and Goals

Despite the remarkable capabilities of large language models (LLMs) such as GPT-5, Gemini, or Claude, their black-box nature often raises concerns about trustworthiness. Explainable AI (XAI) research shows that explanations can improve transparency and trust. Early studies on explanations introduced various explanation provision strategies, such as automatic (always displayed), user-invoked (on-demand), and intelligent (displayed when considered necessary by the system). Among these, intelligent explanations are particularly promising because they reduce information overload while offering necessary insights without extra user effort. Recent work on quantifying uncertainty in language model answers (Chen & Mueller, 2024) offers a promising basis for deciding *when* explanations should be shown. For example, in the context of LLM-based data analysis assistants, an uncertainty quantification could be derived by combining different signals, such as contextual data quality checks and consistency checks across model responses.

The goal of this thesis is to implement such an uncertainty-triggered explanation mechanism that provides explanations only when the system deems it necessary, i.e., detects a higher likelihood of problematic or unreliable output. Basic LLM data assistant prototypes are available, and the code can be provided to the student as a starting point. The overall development process should follow the design science research approach (Hevner et al., 2004) and include a small-scale evaluation of the prototype, e.g., with fellow students.

Required Skills

- Strong interest in (generative) AI and LLMs
- Good English language skills
- Basic programming skills (e.g., Python)

Starting Literature (Topic)

Chen, J., & Mueller, J. (2024). Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5186–5200).

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly, 23*(4), 497.

Sîrbu, A.-M., Schelhorn, T. C., Gnewuch, U. (2025) “Explanation Provision Strategies in LLM-based Data Assistants: Impact on Extraneous Cognitive Load, Trust, and Task Performance,” in Proceedings of the 33rd European Conference on Information Systems (ECIS 2025).

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33.

Starting Literature (Method)

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. *Design science research*. *Cases*, 1-13. https://doi.org/10.1007/978-3-030-46781-4_1

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105. https://doi.org/10.2307/25148625