Nils Feldhus

German Research Center for Artificial Intelligence (DFKI)
Alt-Moabit 91c
10559 Berlin
Berlin, Germany

nils.feldhus@dfki.de
https://nfelnlp.github.io/

1 Research interests

My main research interest is **human-centric explainability**, i.e., making language models more interpretable by building applications that lower the barrier of entry to explanations. I am enthusiastic about **interactive systems** that pique the interest of more people beyond just the experts to learn about the **inner workings of language models**. My hypothesis is that users of language model applications and dialogue systems are more satisfied and trusting if they can look behind the curtain and get easy access to explanations of their behavior.

1.1 Dialogue-based explainability

Human-centered XAI is concerned with incorporating insights from Human-Computer Interaction (HCI) into the field of XAI (Miller, 2019; Ehsan and Riedl, 2020; Weld and Bansal, 2019). Many XAI systems have interactive components, elaborate user interfaces and are evaluated with user studies (Chromik and Butz, 2021; Bertrand et al., 2023). Only recently, however, there has been a push towards conceptualizing dialogue-based XAI systems. Lakkaraju et al. (2022) proposed four modules which are necessary for explanatory conversational systems: Natural language understanding (NLU), explanation algorithm, response generation, and a graphical user interface. Representative systems like TalkToModel (Slack et al., 2023), ConvXAI (Shen et al., 2023), InterroLang (Feldhus et al., 2023), and LLMCheckup (Wang et al., 2024) all implement these four modules.

However, the current conversational XAI systems exhibit a lack of understanding the user and responding to them. This is because they do not consider context and often resemble question answering setups (request and provide explanations). They lack a dedicated dialogue management, as traits of information-seeking (Stepin et al., 2024), mixed-initiative (or proactive) dialogues (Deng et al., 2023), argumentation dialogues (Bex and Walton, 2016) and teacher-student (or tutorial) dialogues (Wachsmuth and Alshomary, 2022; Lee et al., 2023; Liu et al., 2024b) are necessary for a natural explanatory dialogue.

Current research in computational argumentation (Bex and Walton, 2016; Madumal et al., 2019) provides valuable insights into explanatory dialogue interactions, yet it remains relatively abstract and does not cover the full range of explanation moves. Similarly, while didactics literature (Wachsmuth and Alshomary, 2022; Hennessy et al., 2016) defines many moves, it lacks a comprehensive dialogue strategy.

I am currently working on a concept for an explanatory dialogue management which is able to take context into account and easily adapt to user needs. I conduct user studies to examine if LLM-generated explanations are able to take dialogue context into account and, at the same time, beat conventional template-based answers in terms of likeability and perceived faithfulness.

LLMs are getting increasingly better at synthesizing natural language explanations (Wiegreffe et al., 2022) and offer the possibility to hold conversations in various styles, e.g. concise vs. elaborate explanations (Liu et al., 2024a). On top of that, they have been shown to perform dialogue state tracking exceptionally well (Heck et al., 2023). However, LLMs also introduce issues with ground truth, which recent work has started to analyze with test suites (Atanasova et al., 2023) and user studies (Si et al., 2024). I intend to answer the question of whether the faithfulness as perceived by the user matches the actual faithfulness as measured by explanation evaluation and LLM factuality evaluation methods.

1.2 Explanations in tutoring systems

Explanations can also be framed as instructions, e.g. in didactics, where a teacher instructs a student on a concept or topic (Wachsmuth and Alshomary, 2022). Didactics research often debates which teaching strategies lead to the best learning outcome (Roelle et al., 2015). I am investigating if language models can reliably detect if a teacher follows good practices as defined by teaching strategies Feldhus et al. (2024). It turns out that this requires a very thorough definition of acts and high expertise of annotators to achieve a sufficient agreement and trustworthy evaluation results.

A language model that can extract explanation and

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 1–4
September 16–17, 2024.

teaching moves would be helpful for didacticians to self-check and scale up assessments. This is why I am also looking into evaluation measures for generated text, specifically those for measuring how close teachers stick to lesson planning (Feldhus et al., 2024) and accessibility such as readability (Hsu et al., 2024).

Several works have pointed out the difficult of using neural language models for the purpose of tutoring (Macina et al., 2023; Wang and Demszky, 2023). A final goal would be personalized tutoring chatbots that are aware of the user's personality and can adapt their explanatory processes to the expertise and mental model of the user (Fernau et al., 2022).

2 Spoken dialogue system (SDS) research

I believe that SDS research play a vital role in many domains, such as medicine (clinical decision support) and journalism (fact checking). Assistants have a growing presence in our everyday lives and they need to be trustworthy and accountable. Faithful explanations that are grounded in the data, architecture and documentation of the models need to accompany dialogue systems for that reason.

In the coming years, SDS research needs a higher focus on user studies and human evaluation rather than architectures, scaling and exuberant claims of emergent capabilities or agency. With a focus on evaluation and the collection of valuable resources for the growing range of downstream tasks and with the purpose of filling pressing gaps in a multilingual landscape, we can mitigate the actual and present risks for society from uncontrolled systems that already extrude falsehoods and augment harmful biases.

3 Suggested topics for discussion

- How should we design effective explanatory dialogue and conversational XAI systems?
  - Under which circumstances can they depend on LLMs?
  - What findings from other disciplines such as didactics and argumentation should we take into account when building such systems?
- How can the quality of explanation dialogues be evaluated?
- Are LLMs reliable and trustworthy tutoring systems?

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 283–294. https://doi.org/10.18653/v1/2023.acl-short.25.

Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23. https://doi.org/10.1145/3544548.3581314.

Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* 7(1):55–68. https://doi.org/10.3233/AAC-160001.

Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*. Springer, pages 619–640. https://doi.org/10.1007/978-3-030-85616-8_36.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 10602–10621. https://doi.org/10.18653/v1/2023.findings-emnlp.711.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*. Springer International Publishing, Cham, pages 449–466. https://doi.org/10.1007/978-3-030-60117-1_33.

Nils Feldhus, Aliki Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. Association for Computing Machinery, New York, NY, USA, GoodIT '24, page 225–230. https://doi.org/10.1145/3677525.3678665.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets

through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 5399–5421. https://doi.org/10.18653/v1/2023.findings-emnlp.359.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 135–145. https://doi.org/10.18653/v1/2022.sigdial-1.15.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 936–950. https://doi.org/10.18653/v1/2023.acl-short.81.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction* 9:16–44. https://doi.org/https://doi.org/10.1016/j.lcsi.2015.12.001.

Yi-Sheng Hsu, Nils Feldhus, and Sherzod Hakimov. 2024. Free-text rationale generation under readability level control. *arXiv* abs/2407.01384. https://arxiv.org/abs/2407.01384.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *HCAI @ NeurIPS 2022* https://arxiv.org/abs/2202.01875.

Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23. https://doi.org/10.1145/3544548.3581369.

Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024a. TOAD: Task-oriented automatic dialogs with diverse response styles. *To appear in ACL 2024 Findings* https://arxiv.org/abs/2402.10137.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F. Chen. 2024b. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv* abs/2404.03429. https://arxiv.org/abs/2404.03429.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, pages 2357–2372. https://aclanthology.org/2023.eacl-main.173.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, page 1033–1041. https://dl.acm.org/doi/abs/10.5555/3306127.3331801.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007.

Julian Roelle, Claudia Müller, Detlev Roelle, and Kirsten Berthold. 2015. Learning from instructional explanations: Effects of prompts based on the active-constructive-interactive framework. *PLOS ONE* 10(4):e0124115. https://doi.org/10.1371/journal.pone.0124115.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, CSCW '23 Companion, page 384–387. https://doi.org/10.1145/3584931.3607492.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, pages 1459–1474. https://aclanthology.org/2024.naacl-long.81.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learn-

ing models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-023-00692-8.

Ilia Stepin, Katarzyna Budzynska, Alejandro Catalá, Martin Pereira-Fariña, and Jose Maria Alonso-Moral. 2024. Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics. *Argument & Computation* https://content.iospress.com/articles/argument-and-computation/aac220011.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so I could understand": A dialogue corpus for learning to construct explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pages 344–354. https://aclanthology.org/2022.coling-1.27.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Mexico City, Mexico, pages 89–104. https://aclanthology.org/2024.hcinlp-1.9.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, pages 626–667. https://doi.org/10.18653/v1/2023.bea-1.53.

Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62(6):70–79. https://doi.org/10.1145/3282486.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 632–658. https://doi.org/10.18653/v1/2022.naacl-main.47.

Biographical sketch

Nils Feldhus is a final-year PhD student at DFKI under the supervision of Prof. Dr.-Ing. Sebastian Möller at TU Berlin. His research focus is making language model explanations accessible to more target groups such as domain experts and NLP beginners. He presented his work at various conferences, including EMNLP (2021 & 2023), SIGDIAL (2022), ACL (2023), and IJCAI (2022). He holds degrees in computational linguistics (BA) from Heidelberg University and cognitive systems (MSc) from Potsdam University. He is an area chair for ACL Rolling Review since February 2024 for the Interpretability and Analysis of NLP Models track. In his free time, he enjoys music production, cycling, board games, and nature photography.