Patrícia Schmidtová

Charles University
Institute of Formal and Applied Linguistics
Prague, Czech Republic

schmidtova@ufal.mff.cuni.cz
ufal.mff.cuni.cz/patricia-schmidtova

1 Research interests

My research interests lie in the area of **natural language generation** (NLG), more specifically, I focus on the **faithfulness** of NLG. Following Maynez et al. (2020), we define faithfulness as adherence to a given set of inputs (such as a result of a database lookup or a system action). These inputs can either be given by the user with the goal of a given transformation (e.g. data-to-text generation or summarization), or by a dialogue system to compose a reply to the user.

In contrast to numerous works that focus on factuality, i.e. the real-world truth value of a statement, (Azaria and Mitchell, 2023; Lin et al., 2022), I believe that faithfulness is a more useful quality in the realm of **dialog systems** since it measures whether the user received the information they asked for. Thus, my research is directly applicable to the task of **dialog response generation**.

My research is guided by two research questions:

1. How can we determine if a generated text is faithful to its source data?
2. Which factors affect the faithfulness of an LLM's output and how can we manipulate them to achieve better accuracy?

In the following sections, I will outline my progress and plans for how to evaluate faithfulness and thus answer the first research question (Sec 1.1), my plans to understand and improve the faithfulness of systems to seek answers to the second research question (Sec 1.2), and my previous work on treating script generation as a dialogue system task (Sec 1.3).

1.1 Evaluation of faithfulness

There are several challenges when designing a robust protocol to evaluate faithfulness. Most metrics rely on the presence of gold reference data (Papineni et al., 2002; Zhang et al., 2020; Kane et al., 2020) that is not always available. Furthermore, some problems have more than one correct solution, and comparing to an arbitrary reference might not always favor the best outputs.

Additionally, in our work examining **data contamination** (i.e. presence of testing data in the training data) (Balloccu et al., 2024), we found that many datasets with gold annotations, such as several variants of MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Ye et al., 2022) and some datasets used for DSTC (Zhao et al., 2023) were leaked to closed-source language models by users. With works examining the presence of datasets in CommonCrawl (Li et al., 2024), we cannot even be entirely sure that open-weight models with secret training data, such as Mistral (Jiang et al., 2023) or Llama2 (Touvron et al., 2023) are free from data contamination. This casts a shadow of doubt on whether the models truly generalize well or whether a part of their success is due to data contamination.

Therefore, in my research, I will focus on reference-free evaluation methods that can be used on freshly mined data, such as the QUINTD dataset (Kasner and Dušek, 2024). We have seen some success using LLMs as evaluators for dialogue response generation (Plátek et al., 2023) and we are currently extending this work on new datasets and with comparison to crowd-workers of several proficiency levels determined based on a qualification screening test. Currently, there are also reference-free metrics based on **natural language inference** (NLI), however, they are not yet equipped to deal with structured data or with data of various lengths. We intend to address this issue in our future work.

We do not intend to replace human evaluation using these methods since insights gained by a well-performed human analysis are unparalleled. We rather see automatic evaluation as a proxy for situations where time and resources are limited, such as in a development cycle when trying to estimate the effect of a change. Additionally, human and automatic evaluation should complement each other to assess the strengths and weaknesses of a system comprehensively.

To simplify human (or LLM) annotation of LLM faithfulness errors, my colleagues and I developed a tool called factgenie[1] (Kasner et al., 2024), which will be presented at INLG as a demo the week after YRRSDS. Finally, we have prepared a comprehensive survey of how automatic evaluation is generally performed in NLG and extended a set of best practices (Schmidtová et al., 2024). This work will also be presented at INLG. One of our

[1] https://github.com/kasnerz/factgenie

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 21–24

main findings was that evaluation in NLG is currently very divided. The most prominently used metrics are based on N-gram overlap, such as BLEU (Papineni et al., 2002), which is unfortunate, since Reiter (2018) shows that they have little informational value in NLG.

1.2 Understanding and improving faithfulness

When trying to understand the faithfulness of LLMs to a given input, prompts are the easiest external factor to examine. Axelsson and Skantze (2023) observed that asking an LLM to stick to the provided facts indeed increases their faithfulness. In our research, we intend to explore how various circumstances, such as prompt length, grammatical correctness, or the presence of specific instructions, affect the faithfulness of a language model.

Moreover, we also intend to use **probing** to observe how the different prompts activate different parts of the network and thus elicit different results. We draw inspiration from work where probing was used to seek out and modify facts stored in LLMs' trained weights (Meng et al., 2022) or to classify if an LLM believes that a statement supplied by the user on the input is true (Azaria and Mitchell, 2023).

1.3 Previous work on theatre play script generation

The majority of my past work on theatre play generation was performed with a single language model predicting the next character utterance (Schmidtová et al., 2022). However, one of the downsides of this approach was the lack of consistency in the characters' personalities. As a small side project, we decided to treat this task as a conversation between three language models, each of them fine-tuned to represent a separate character (Schmidtová et al., 2022). To keep things simple, we classified characters in movie scripts into pessimists, optimists, and realists by observing the average sentiment of their utterances. We showed that by training each model separately, the consistency of characters was indeed improved.

2 Spoken dialogue system (SDS) research

The arrival of large language models trained using reinforcement learning from human feedback changed the way how the public perceives dialogue systems and what to expect from them. I believe there are two directions in research we should pay attention to in the next 5-10 years:

Multidisciplinary collaboration We might not even be aware of all the ways how the public uses dialogue systems and often only hear about the use cases where something went wrong, such as a lawyer citing non-existent cases (Merken, 2023). I believe it is important to connect with other fields, especially psychology, to have a better understanding of how SDSs impact the users so we can make more informed decisions about how we design and present the systems to make them safer.

Educating the public Last, but not least, we see many public figures make bold statements about how LLMs will make entire careers, such as programmers, obsolete. Generally, the boldest claims do not come from researchers, but rather from executives seeking to increase profits of the companies they run. For this reason, I believe that communicating research to the public has grown equally as important as the research itself. Scientists should be the figures that the public looks to with questions, yet they are often not very visible outside of academic grounds. As young researchers, we can start small, for example by giving talks to high school students or interested communities around us.

3 Suggested topics for discussion

These are the topics I would like to suggest for discussion:

- **Data contamination**: to what extent should we examine and worry about dialogue datasets being contained in CommonCrawl or the training sets of closed-source models?
- **Evaluation**: should we strive for a more general and unified set of evaluation practices or rather try to adapt the metrics used to the presented dialogue system?
- **Multidisciplinary collaboration**: Other fields, such as robotics or social sciences can be very beneficial to SDS and provide insights to make them better and safer. On the other hand, the structure and funding distribution of universities does not always favor such collaboration. How do others tackle this, if at all?

Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 40222 and SVV 260 698.

References

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs. In Albert Gatt, Claire Gardent, Liam Cripwell, Anya Belz, Claudia Borg, Aykut Erdem, and Erkut Erdem, editors, *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*. Association for Computational Linguistics, Prague, Czech Republic, pages 39–54. https://aclanthology.org/2023.mmnlg-1.5.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, pages 67–93. https://aclanthology.org/2024.eacl-long.5.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 5016–5026. https://doi.org/10.18653/v1/D18-1547.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 422–428. https://aclanthology.org/2020.lrec-1.53.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*. Association for Computational Linguistics, Online (Dublin, Ireland), pages 28–37. https://aclanthology.org/2020.evalnlgeval-1.4.

Zdeněk Kasner and Ondřej Dušek. 2024. Beyond reference-based metrics: Analyzing behaviors of open llms on data-to-text generation.

Zdeněk Kasner, Ondřej Plátek, Patrícia Schmidtová, Simone Balloccu, and Ondřej Dušek. 2024. factgenie: A framework for span-based evaluation of generated texts. https://arxiv.org/abs/2407.17863.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. An open source data contamination report for large language models. https://arxiv.org/abs/2310.17589.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=-h6WAS6eE4.

Sara Merken. 2023. New york lawyers sanctioned for using fake chatgpt cases in legal brief. https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Ondřej Plátek, Vojtech Hudecek, Patricia Schmidtova, Mateusz Lango, and Ondrej Dusek. 2023. Three ways of using large language models to evaluate chat. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Pro-*

ceedings of The Eleventh Dialog System Technology Challenge. Association for Computational Linguistics, Prague, Czech Republic, pages 113–122. https://aclanthology.org/2023.dstc-1.14.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44(3):393–401. https://doi.org/10.1162/coli-a-00322.

Patrícia Schmidtová, Rudolf Rosa, David Košťák, Tomáš Studeník, Daniel Hrbek, Tomáš Musil, Josef Doležal, Ondřej Dušek, David Mareček, Klára Vosecká, Máša Nováková, Petr Žabka, Alisa Zakhtarenko, Dominik Jurko, Martina Kinská, Tom Kocmi, and Ondřej Bojar. 2022. *THEaiTRE: Generating Theatre Play Scripts using Artificial Intelligence*. Institute of Formal and Applied Linguistics, Prague, Czechia.

Patrícia Schmidtová, Dávid Javorský, Christián Mikláš, Tomáš Musil, Rudolf Rosa, and Ondřej Dušek. 2022. Dialoguescript: Using dialogue agents to produce a script. https://arxiv.org/abs/2206.08425.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. To appear at INLG 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 351–360. https://doi.org/10.18653/v1/2022.sigdial-1.34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge.

Biographical sketch

Patrícia Schmidtová is a second-year PhD student at Charles University advised by Ondřej Dušek. Currently, she is researching how to comprehensively and reliably evaluate large language models, especially the faithfulness of their outputs to the data they are given. She plans to explore why LLMs respond to some prompts better than others by using interpretability techniques.

Before her PhD, she was a member of the THEaiTRE research team which succeeded in producing the world's first AI-scripted theater play. She also has six years of industry experience in NLP application development, mostly working on devising components for robotic process automation (RPA).