Kristýna Onderková

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

onderkova@ufal.mff.cuni.cz
ufal.mff.cuni.cz/kristyna-onderkova

1 Research Overview

My research focuses on improving **textual inference in large language models (LLMs) for natural language generation (NLG)**. LLMs are increasingly used to generate reports, summaries, insights, and answer questions about data. However, their outputs are often factually inaccurate (Thomson et al., 2023; Islam et al., 2023), have difficulties with inference operations (Creswell et al., 2023; Wu et al., 2024) and produce shallow outputs as most benchmarks do not require much inference (e.g. WebNLG (Gardent et al., 2017)) and comprehensive content selection (Puduppully et al., 2019). This limits their usefulness in contexts such as communicating complex information and decision making, where people need meaningful and reliable outputs.

I am particularly interested in data-to-text generation, which requires both **faithfulness** to the provided data and an intuition for what might be **interesting** and **important**. For example, rather than stating superficially "*The tornado caused 12 fatalities in Missouri*", deeper inference could be "*Missouri had the highest number of fatalities, nearly a third of all deaths*". Such tasks are often under-specified, forcing both models and humans to make implicit presuppositions, which can cause errors when they differ. Beyond prompting LLMs, I want to integrate them with symbolic operations through code generation to make inferences over the data, ensuring that the outputs are more faithful and interpretable.

My work is driven by the following research questions:

1. How can we use LLMs coding abilities to make more faithful and deeper inferences about the data using neuro-symbolic systems?

2. How do implicit presuppositions about the data influence both LLMs' outputs and their evaluation?

3. How do human's implicit presuppositions influence which insights are perceived as meaningful and interesting?

These questions are general across tasks and domains, and I aim to explore them as such. Now I focus on table-to-text generation across domains, choosing the science reporting domain as an illustration.

The following sections outline my progress and plans regarding the first research question (1.1), the presupposition questions (1.2), and my previous related work on abductive reasoning (1.3).

1.1 Table-to-text generation

Table-to-text generation (Liu et al., 2018; Parikh et al., 2020), a subtask of data-to-text generation, has been addressed by both rule-based (Gatt and Krahmer, 2018) and neural methods (Nan et al., 2022; Zhao et al., 2023a). Inspired by code generation LLMs and question-answering approaches using SQL (Cheng et al., 2023; Kong et al., 2024), I developed a LLMs pipeline that generates ideas for insights, then generates SQL to symbolically retrieve relevant information and formulate the final insight. However, this approach did not substantially surpass the baseline, largely due to messy input tables. I will extend it by creating knowledge bases enhanced with generated presuppositions. Moreover, since current benchmarks involve little inference, generated insights are often obvious (e.g., January comes before February) and lack diversity. I intend to use SPARQL for the retrieval and Prolog/Isabella for the inferences on the knowledge base to generate deeper insights, building on "hard critique" in natural language explanations (Quan et al., 2025, 2024). This approach might require fine-tuning, as the setting is less common than text-to-SQL.

For experiments, I used LogicNLG and LoTNLG (Zhao et al., 2023b; Chen et al., 2020) benchmarks, which are already known to LLMs, raising concerns of data contamination (Oren et al., 2024; Xu et al., 2024). To address this, I created FreshTab, a dynamic live benchmark accepted to INLG 2025. FreshTab collects table-to-text datasets from Wikipedia using only tables added after a specified knowledge cutoff date. It also includes a basic domain distribution (sport, politics, culture, mix) and supports non-English Wikipedias. Automatic metrics (e.g., TAPEX (Liu et al., 2022)) indicated that new and more diverse data are more difficult for LLMs, while human evaluation suggests that it is the trained metrics that struggle more with memorization.

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 17–20

1.2 Presuppositions

During my work on the table-to-text generation, I observed that the task is often under-specified, especially with real-world, messy data. Presuppositions about the table structure are not always met (e.g., multiple different units in one column, contain sub-tables, are "transposed", or graphical). Preliminary observations suggest that additional presuppositions arise also from: (i) expertize (e.g., *fresh breeze* means exactly 29–38 km/h to meteorologists but not to most people), (ii) cultural context (e.g., rating scales - is lower better or worse?), (iii) temporal changes (e.g., changes in rules of a sport game) and others.

The aim is to investigate such presuppositions in the context of NLG and its evaluation, building on Han et al. (2023)'s work in question answering. The presuppositions would help in several ways: (i) understanding the structure of messy input data, (ii) personalizing the generation based on users' own presuppositions to make them more interesting, and (iii) making the evaluation of faithfulness more rigorous by stating the presuppositions.

Firstly, I need to conduct a more thorough literature review across additional relevant research directions. Then, I will analyze existing data (e.g. tables with corresponding insights from Wikipedia or news articles), along with human studies, to evaluate how expertise and the complexity of logical operations affect perceived interestingness and relevance of insights. I also want to symbolically evaluate the generated insights against the extracted knowledge bases and presuppositions.

1.3 Previous work on abductive reasoning in NLG

During my master's thesis, I investigated how language models perform on abductive reasoning tasks using the ART dataset (Bhagavatula et al., 2020). Although humans use abductive reasoning daily, it is used in automated planning and is effective at handling incomplete information, it has been relatively underexplored. As abductive reasoning infers causes for observed effects based on both explicit and presupposed knowledge, it highlights the importance of implicit presuppositions in inference.

Building on my industry experience with the classical NLG pipeline (Reiter and Dale, 1997), I studied the applicability of prompting language models (*Flan-T5*, *Falcon Instruct*) with abductive tasks for content selection and discourse planning. Experiments showed some effectiveness in content selection (mapping data and rules to outcomes such as snow, rain, or clear weather), but discourse planning proved to be more challenging (Onderková and Nickles, 2023).

2 Short Biography

Kristýna Onderková is a first-year PhD student at Charles University, supervised by Ondřej Dušek. Her research focuses on improving textual inference and reasoning in language models, particularly in data-to-text generation. She works on neuro-symbolic approaches that combine statistical models with symbolic operations to improve content selection, factuality, interpretability, and controllability. She is interested in how language models can uncover meaningful insights from data. She plans to explore how language models presuppose information about data, aiming to make these presuppositions explicit to reduce under-specification and improve inference.

She has an interdisciplinary background with a master's degree in both Computer Science and Physics. She worked for four years in industry on data analytics, machine learning, and natural language processing, including NLG for news reporting (e.g. weather, sports, elections) and LLM-based agent applications.

3 Questions for Senior Researchers

I respectfully ask the senior researchers for their insight on the following questions:

- How can researchers select research ideas that are likely to make substantial contributions and avoid producing short-lived papers?
- How do you design NLG experiments to test hypotheses while identifying and controlling for confounding factors rigorously?
- How can experiments with language models be designed to account for prompt variability and ensure robust findings?
- How can reasoning in NLG outputs be effectively evaluated, distinguishing genuine reasoning from simple retrieval?
- What strategies help researchers keep up with new publications and emerging methods?

4 Topics for Roundtable Discussions

I would like to propose the following discussion topics:

- **Presuppositions:** How do implicit presuppositions in under-specified tasks affect model behavior and evaluation, and would explicitly stating them improve interpretability and reliability?
- **Evaluation of usefulness:** How can we ensure that the metrics we use and the tasks we address correspond to outcomes that are meaningful and relevant to humans?
- **AI-assisted research practices:** In what ways can artifical inteligence tools be effectively integrated into research workflows, such as literature review, coding, and writing assistance, while maintaining scientific rigor and critical thinking?

Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303), the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO, and Charles University project SVV 260 698.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen Tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404* .

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding Language Models in Symbolic Languages. Kigali, Rwanda. https://openreview.net/forum?id=lH1PV42cbF.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=3Pf3Wg6o-A4.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *10th International Conference on Natural Language Generation*. ACL Anthology, pages 124–133.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation 61(1):65–170.

Wookje Han, Jinsol Park, and Kyungjae Lee. 2023. Prewome: Exploiting presuppositions as working memory for long form question answering. In *2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*. Association for Computational Linguistics (ACL), pages 8312–8322.

Saad Obaid Ul Islam, Iza Škrjanec, Ondřej Dušek, and Vera Demberg. 2023. Tackling hallucinations in neural chart summarization. In *Proceedings of the 16th international natural language generation conference*. pages 414–423.

Kezhi Kong, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Chuan Lei, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2024. Opentab: Advancing large language models as open-domain table reasoners. In *12th International Conference on Learning Representations, ICLR 2024*.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*. https://openreview.net/forum?id=O50443AsCP.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-Text Generation by Structure-Aware Seq2seq Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 32. Number: 1. https://doi.org/10.1609/aaai.v32i1.11925.

Linyong Nan, Lorenzo Jaime Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou, and Dragomir Radev. 2022. R2d2: Robust data-to-text with replacement detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. pages 6903–6917.

Kristýna Onderková and Matthias Nickles. 2023. Exploring abductive reasoning in language models for data-to-text generation. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*. IEEE, pages 1–4.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving Test Set Contamination in Black-Box Language Models. In *ICLR*. Vienna, Austria. https://openreview.net/forum?id=KS8mIvetg2.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1173–1186.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*. volume 33, pages 6908–6915.

Xin Quan, Marco Valentino, Danilo S Carvalho, Dhairya Dalal, and André Freitas. 2025. Peirce: Unifying material and formal reasoning via llm-driven neuro-symbolic refinement. *arXiv preprint arXiv:2504.04110* .

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. 2024. Verification and refinement of natural language explanations through LLM-symbolic theorem proving. In Yaser Al-Onaizan, Mohit Bansal,

and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, pages 2933–2958. https://doi.org/10.18653/v1/2024.emnlp-main.172.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural language engineering* 3(1):57–87.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language* 80:101482.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pages 1819–1862.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244* .

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Yu Flores, and Dragomir Radev. 2023a. Loft: Enhancing faithfulness and diversity for table-to-text generation via logic form control. *arXiv preprint arXiv:2302.02962* .

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. pages 160–175.