YNLG 2025

The 1st Workshop for Young Researchers in Natural Language Generation

Proceedings of the Workshop

October 29, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-328-9

Preface

The main objective of the workshop is to create a platform where young researchers can receive feedback on their research ideas, meet like-minded participants, discuss current trends, and find collaborators for future work. In the long term, this effort aims to create a welcoming network that attracts NLG researchers who are not yet members of the SIGGEN community. Ideally, this effort will organically grow the SIGGEN community from the ground up.

We take inspiration from the Young Researchers' Roundtable on Spoken Dialogue Systems (YRRSDS):[1] the workshop for young researchers working on dialogue systems. It has been running for over 20 years, helped make the SIGDIAL community more welcoming, and created the space for many collaborations.

The inaugural Young Researchers in Natural Language Generation Workshop (YNLG) was held in conjunction with the 18th International Natural Language Generation Conference (INLG) in Hanoi, Vietnam from October 29 to November 2, 2025.

Young researchers submitted a 2-page position paper reporting on their past, present, and future work, a short bio, and topic suggestions for discussions. The participants then presented their submissions during a poster session, which enabled them to gather feedback from a wider audience. Each submission was carefully reviewed by three senior researchers from our Program Committee. We extend our deep gratitude to the Program Committee members for their insightful reviews; their contributions have been invaluable in offering critical feedback to the participants at this pivotal stage in their careers.

YNLG accepted all 8 submissions received, with 4 opting for their work to be archival. Alongside the oral sessions and roundtables, the program featured two outstanding keynote presentations. We would like to express our gratitude and acknowledge our keynote speakers: David M. Howcroft (University of Aberdeen) and Hadas Kotek (Apple) for their inspiring talks full of advice for young researchers. We further encouraged follow-up discussions by hosting a panel discussion in the afternoon, including the two keynote speakers as well as Guanyi Chen (Central China Normal University) and Saad Mahamood (Shopware). Finally, the evening program included a dinner fostering casual interactions between participants and organizers.

We extend our gratitude to the INLG organizers for making sure the conference ran seamlessly and was enjoyed by all attendees.

On behalf of the YNLG 2025 Organizers,
Patrícia Schmidtová and Nils Feldhus

[1] https://sites.google.com/view/yrrsds2024/program

Organizing Committee

1. Alyssa Allen (The Ohio State University)
2. Nils Feldhus (TU Berlin, BIFOLD)
3. Rudali Huidrom (Dublin City University, ADAPT)
4. Michela Lorandi (Dublin City University, ADAPT)
5. Adarsa Sivaprasad (University of Aberdeen)
6. Patricia Schmidtova (Charles University)

Program Committee

1. Ekaterina Artemova (Toloka)
2. Simone Balloccu (TU Darmstadt)
3. Nikolay Bogoychev (Meta)
4. Ondřej Dušek (Charles University)
5. Sebastian Gehrmann (Bloomberg)
6. Jindřich Helcl (University of Oslo)
7. David M. Howcroft (University of Aberdeen)
8. Tatsuya Ishigaki (AIST)
9. Zdeněk Kasner (Charles University)
10. Mateusz Lango (Charles University)
11. Jindřich Libovický (Charles University)
12. Tomasz Limisiewicz (University of Washington, Meta)
13. Chenghua Lin (University of Manchester)
14. Saad Mahamood (trivago)
15. Simon Mille (Dublin City University, ADAPT)
16. Shashi Narayan (Google DeepMind)
17. Ehud Reiter (University of Aberdeen)
18. Elizabeth Salesky (Google DeepMind)
19. Fahime Same (trivago)
20. David Schlangen (Potsdam University)
21. Veronika Solopova (TU Berlin)
22. Emiel van Miltenburg (Tilburg University)
23. Jing Yang (TU Berlin)

Table of Contents

Conference Program

Data Generation and Low-resource Scenarios
Tatiana Anikina

Improving Factual Accuracy in Neural Data-to-Text Generation through Input Quality and Scalable Evaluation
Barkavi Sundararajan

Insight discovery in structured data
Kristyna Onderkova

Efficient methods for building LLMs for low-resourced languages
Nalin Kumar

Tatiana Anikina

German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Germany
tatiana.anikina@dfki.de
www.anikina.de

1 Research Interests

As a computational linguist I have been working on a variety of natural language processing (NLP) topics including coreference resolution, dialogue processing, explainability, and synthetic data generation. My earlier research includes e.g. the topic of **coreference resolution** in Abstract Meaning Representation which was based on my Master's thesis (Anikina et al., 2020). I have also participated in multiple shared tasks on coreference resolution in dialogue (Anikina et al., 2021, 2022; Skachkova et al., 2023), dialogical argument mining (Binder et al., 2024), and multilingual claim normalization (Anikina et al., 2025b).

My current research focuses mostly on efficient data augmentation strategies and low-resource scenarios. For instance, in my submission to the Student Research Workshop at ACL (Anikina, 2023) I focused on the task of adapting NLP models to **dialogue processing in the emergency response domain**. Having a system for dialogue intent classification and slot tagging that is robust, flexible, and data efficient is a challenging task, especially given that the emergency response domain must cover a wide variety of scenarios (e.g. fires, floods, building collapse etc.) and the available annotations are typically very scarce. In this work I evaluated different approaches towards using adapters with Transformer models and applied different **data augmentation** techniques in a **low-resource setting**, e.g. backtranslation and token replacements as well as contextual augmentation with speaker information, previous turns and dialogue summary.

Another recent work that has been recently accepted at EMNLP (Anikina et al., 2025a) compares various generation strategies for producing synthetic data in low-resource languages. Here we performed a systematic evaluation of demonstrations in different languages, label-based summaries, and self-revision that can be used for **synthetic data generation**, and evaluated the performance of different prompting strategies across 11 typologically diverse languages, including several extremely low-resource ones such as Welsh and Azerbaijani. This work shows that it is possible to achieve very good performance when fine-tuning models on the synthetic data but such data should be carefully selected and ideally revised by language models.

Currently, I am doing a research internship in the JetBrains AI Writing Assistance team, preparing a dataset with synthetic data for training a commit completion model. The goal of this project is to find an optimal way of **generating new good quality data for commit messages** that describe the code changes. Synthetic data are needed here since the existing messages collected from the sources like *GitHub* are often noisy, ungrammatical or they may not fully reflect all the relevant changes (e.g. "fix the bug" is not a good commit message compared to "fix the logic of error handling in ProjectClass"). Ideally, good messages would be fluent, coherent with the commit history and correct with respect to the code diff. Together with the team I have been experimenting with various models and approaches, including code diff summarization, chain-of-thought prompting (Fu et al., 2022), and different ways of selecting the demonstrations.

2 Data Generation and Low-resource Scenarios

Modern Large language models (LLMs) such as *ChatGPT*, *Llama*, *Qwen*, *Gemini*, and many others demonstrate impressive capabilities of producing natural language text that is fluent and typically follows the user instruction encoded in a prompt. Such LLMs are often used to generate new data for training smaller downstream models in an efficient way, they can be used for both data augmentation and generation (Dai et al., 2023; Piedboeuf and Langlais, 2023; Cegin et al., 2025). LLM-generated texts can be useful for a variety of tasks from intent recognition (Cegin et al., 2024) and news classification (Piedboeuf and Langlais, 2023) to hate speech detection (Sen et al., 2023) and sentiment analysis (Onan, 2023). Increasingly more complex tasks can be tackled by LLMs, Pranida et al. (2025) compared the data that were generated by humans, with machine translation, and LLM-generated texts for culturally nuanced commonsense reasoning in low-resource languages and found that LLM-generated data were more beneficial for downstream classification than machine translation.

However, LLMs are mostly trained on the English data and even though some multilingual versions exist, e.g. *Aya* or *Gemma 3*, they do not cover many languages, es-

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 1–3

pecially the low-resourced ones. Although some works explore generation also for low-resource languages (Zotova et al., 2021; Chung et al., 2025) the general trend is to expect lower quality of the synthetic data for those languages that were unseen (or barely seen) during training and more research needs to be done on finding the optimal generation strategies for such languages, potentially combining cross-lingual transfer, demonstration-based prompting, and LLM revision, e.g. expanding on the optimal strategies identified in (Anikina et al., 2025a).

Another interesting and underexplored venue of research is the low-resource setting with respect to the task, not necessarily the language. For instance, Ben Abacha et al. (2023) work towards summarization of doctor-patient conversations in a clinical setting. They found that the generated summaries show high fluency score, but still struggle with hallucinations and miss 33% important medical facts. Although using synthetic data like automatically generated summaries of doctor-patient conversations is very attractive as it can reduce the documentation burden for medical professionals, there is still a large room for improvement. Another low-resource domain where synthetic data could be especially valuable is emergency response. Privacy constraints and the wide variety of possible scenarios make real data scarce, yet dialogue between first responders still needs to be summarized and classified. To enable this, LLMs must first be trained on synthetic datasets that capture diverse emergency situations and incorporate the relevant terminology.

3 Questions and Discussion Topics

I would like to suggest the following discussion topics and questions to senior researchers:

- **Evaluation:** What are the best was of evaluating the quality of generated data? Often we do not have the gold standard data that the generated data can be compared to and have to resort to either reference-free metrics or rely solely on the performance of a downstream model that was trained on the generated data. Nowadays researchers also use various LLMs to score the responses, applying techniques like e.g. LLM-as-a-Judge (Wang et al., 2024) but how reliable are these judgments? How can we estimate and avoid potential bias?

- **Controlled Generation:** What are the best practices in generating outputs conditioned on specific knowledge or inputs? In many real life situations there are some initial requirements and constraints on the data that need to be respected during the generation. How can we integrate such constraints and check that they are followed in an efficient way?

- **Synthetic Data Recycling:** An increasing amount of web content is now generated by language models. As new training datasets are collected for the next generation of models, it is inevitable that some of this synthetic content will be included. How problematic is this phenomenon, and what long-term consequences might it have, e.g., could it reduce the diversity and quality of model outputs? To what extent is recycling synthetic data a viable strategy, and when should it be avoided?

Biographical Sketch

I hold an M.Sc. in Language Science and Technology from Saarland University. After graduation, I joined the German Research Center for Artificial Intelligence (DFKI) as a PhD student and researcher. Currently, I am on leave from my PhD program to pursue an internship at JetBrains, where I am part of the AI Writing Assistance team. My research interests include efficient data augmentation, dialogue processing, multilingual and low-resource scenarios.

References

Tatiana Anikina. 2023. Towards efficient dialogue processing in the emergency response domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.

Tatiana Anikina, Ján Cegin, Jakub Simko, and Simon Ostermann. 2025a. A rigorous evaluation of llm data generation strategies for low-resource languages. *ArXiv*, abs/2506.12158.

Tatiana Anikina, Alexander Koller, and Michael Roth. 2020. Predicting coreference in Abstract Meaning Representations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–38, Barcelona, Spain (online). Association for Computational Linguistics.

Tatiana Anikina, Cennet Oguz, Natalia Skachkova, Siyu Tao, Sharmila Upadhyaya, and Ivana Kruijff-Korbayova. 2021. Anaphora resolution in dialogue: Description of the DFKI-TalkingRobots system for the CODI-CRAC 2021 shared-task. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–42, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tatiana Anikina, Natalia Skachkova, Joseph Renner, and Priyansh Trivedi. 2022. Anaphora resolution in dialogue: System description (CODI-CRAC 2022 shared task). In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 15–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tatiana Anikina, Ivan Vykopal, Sebastian Kula, Ravi Kiran Chikkala, Natalia Skachkova, Jing Yang, Veronika Solopova, Vera Schmitt, and Simon Ostermann. 2025b. dfkinit2b at checkthat! 2025: Leveraging llms and ensemble of methods for multilingual claim normalization. In *CLEF 2025 Working Notes. Conference and Labs of the Evaluation Forum (CLEF-2025), Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, September 9-12, Madrid, Spain*. CEUR Workshop Proceedings.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Arne Binder, Tatiana Anikina, Leonhard Hennig, and Simon Ostermann. 2024. DFKI-MLST at DialAM-2024 shared task: System description. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 93–102, Bangkok, Thailand. Association for Computational Linguistics.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Use random selection for now: Investigation of few-shot selection strategies in llm-based text augmentation for classification. *Preprint*, arXiv:2410.10756.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. LLMs vs established text augmentation techniques for classification: When do the benefits outweight the costs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10476–10496, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. *arXiv preprint arXiv:2502.15419*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

Yao Fu, Hao-Chun Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *ArXiv*, abs/2210.00720.

Aytuğ Onan. 2023. Srl-aco: A text augmentation framework based on semantic role labeling and ant colony optimization. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101611.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore. Association for Computational Linguistics.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages. *Preprint*, arXiv:2502.12932.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Natalia Skachkova, Tatiana Anikina, and Anna Mokhova. 2023. Multilingual coreference resolution: Adapt and generate. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 19–33, Singapore. Association for Computational Linguistics.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-taught evaluators. *CoRR*, abs/2408.02666.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications*, 170:114547.

Barkavi Sundararajan

University of Aberdeen
King's College
Aberdeen
United Kingdom

b.sundararajan.21@abdn.ac.uk

1 Research Overview

Neural Language Models have become central to Natural Language Generation (NLG) research and can produce fluent and coherent text (Badaro et al., 2023). However, when models generate text from complex, structured or long-form data such as tables or event logs, they often hallucinate and introduce factual errors (Jacovi et al., 2025; Rebuffel et al., 2021). These hallucinations limit the practical deployment of large language models (LLMs) in applications where factual accuracy is critical. In my research, *factual accuracy* refers to the faithfulness of the generated text to the given input data. My PhD focuses on reducing hallucinations and improving factual accuracy in data-to-text generation, which I address through two core approaches: (i) analysing how input quality and structure improve factual accuracy, and (ii) developing a manual error annotation protocol and extending it into an LLM-as-Judge framework. This work aims to assess when automatic evaluation can complement human annotation and enable larger-scale evaluation.

Past work: In my initial work **(Sundararajan et al., 2022)**, I fine-tuned a T5-base model (Raffel et al., 2023) on the ToTTo dataset (Parikh et al., 2020) using the published configuration and compared its outputs with two GeM benchmark models (Gehrmann et al., 2021), which at the time represented the top-performing ToTTo baselines as ranked by PARENT, BLEURT, and BLEU. Using a manual error annotation protocol adapted from Thomson et al. (2023), extended to include `DATE-DIMENSION` and `NON-ENGLISH` error categories, I analysed more than 3,000 outputs in the politics domain and categorised errors into eight types (`WORD`, `NAME`, `NUMBER`, `DATE-DIMENSION`, `CONTEXT`, `NOTCHECKABLE`, `NON-ENGLISH`, and `OTHER`).

- By capturing factual errors more effectively than surface-level automatic metrics, this protocol *revealed systematic weaknesses such as verb mispredictions, numeric swaps, and hallucinations around dates* and demonstrated limitations that BLEU (Papineni et al., 2002) and similar metrics fail to detect, reflecting broader evaluation challenges discussed in Gehrmann et al. (2022).
- The analysis also highlighted challenges when reasoning over complex tables, especially for tables with more than 20 rows.
- The annotation protocol achieved substantial inter-annotator agreement between two independent annotators (Cohen's Kappa = 0.79) on a 10% subset of the outputs, confirming the reliability of the categories (see Appendix A for the procedure).

Building on this, I investigated whether hallucinations in neural table-to-text models can be traced back to problems in the tabular input in the ToTTo dataset **(Sundararajan et al., 2024)**. I manually annotated 1,837 outputs in the politics domain using a further-adapted protocol that included an `ADDITION` category and excluded `NON-ENGLISH`. On a 6.5% overlapping subset of the outputs (120 in total) annotated independently by three annotators, the annotation achieved substantial inter-annotator agreement (Fleiss' Kappa = 0.622).

- This analysis identified several recurring input problems: *non-atomic cell values, insufficient input, longer table inputs, complex tables*, and *other politics domain-specific* input problems.
- I then systematically corrected the non-standard tabular inputs to a standard form (see Appendix B). Across 210 samples, input correction reduced factual errors by 62% in T5-base and 57% in T5-large.
- I extended this to Llama2 models (Touvron et al., 2023), correcting inputs decreased factual errors by 52% for Llama2–7B and 76% for Llama2–13B.

These results demonstrate that many hallucinations arise from garbage-in, garbage-out effects: poorly structured inputs lead to more factual errors, while corrected inputs substantially improve factual accuracy. This finding is consistent with prior work, which shows that improving inputs in NLG datasets leads to better outputs (Dušek and Kasner, 2020).

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 4–10

Recent work: In this work (accepted at INLG 2025), I investigated how different input structures influence the factual accuracy of LLM-generated NBA game summaries from play-by-play data. This is a challenging text generation task with over 450 rows and 13 columns because play-by-play data contains rich information about plays, including entities (teams and players), events/actions (e.g., jump shot, layup, dunk, free throw, rebound), outcomes (missed or made for any attempt), and attributes (score, time, period, distance). To empirically study the impact of input structuring on model output, we experimented with three input data formats: unstructured (control), row-based, and JSON, and generated 180 summaries using Llama3.1-70B and Qwen2.5-72B (Qwen et al., 2025).

We manually annotated factual errors using a further adapted protocol that distinguishes `WORD-OBJECTIVE` (fact-checkable) from `WORD-SUBJECTIVE` (opinion). Error counts were normalised to 'errors per 100 words'.

- A two-way repeated-measures ANOVA on input structure (unstructured, row-based, JSON) revealed significant main effects on error rates ($p < 0.05$).
- JSON inputs cut error rates by 69% for Llama3.1-70B and 65% for Qwen2.5-72B versus unstructured text. Row-based inputs cut errors by 54% for Llama3.1-70B and 51% for Qwen2.5-72B, and switching from row-based to JSON cuts errors by another 32% for Llama3.1-70B and 27% for Qwen2.5-72B.
- Error-type analysis shows Llama3.1-70B produces more numeric mistakes, whereas Qwen2.5-72B makes more naming and subjective-wording errors. Word objective errors are predominant among both models.

These findings demonstrate that structured input representations substantially improve factual accuracy in complex, event-driven summarisation tasks such as sports reporting, although limitations remain in scaling manual annotations and in generalizing to other domains.

Ongoing and future work: The recent study shows that structured inputs improve factual accuracy but also expose the limits of manual evaluation, which is costly and difficult to scale. To address this, I am developing an LLM-as-Judge framework that operationalises my human error annotation protocol into structured prompts and directs judge LLMs to assess atomic factual claims against play-by-play input logs. I will evaluate agreement with human annotations using accuracy metrics and determine when automatic judgments can reliably complement manual ones. As part of this ongoing work, I plan to extend the automatic evaluation framework to other event-driven sports such as ice hockey to examine its generalizability.

2 Biography

Barkavi Sundararajan is a final-year PhD student in Computer Science at the University of Aberdeen. Her research interests lie in reducing hallucinations and improving factual accuracy in data-to-text generation, with a particular focus on developing reliable models and robust evaluation methods for factuality. She is a part-time Data Scientist intern at the UK Ministry of Justice, where she codifies free-text assessments from large-scale datasets into structured data to support projects such as Safer Streets UK. She plans to pursue a career as a research scientist or applied scientist at the intersection of natural language processing and high-stakes decision-making domains (where factual accuracy is critical).

3 Questions for Senior Researchers

1. What role can knowledge graphs play in improving factuality in neural NLG for long-form inputs, compared with other approaches such as retrieval or input restructuring?
2. How can one best approach the transition from a PhD in NLG to Research Scientist or Applied Scientist roles in industry labs?
3. As a PhD student, how can I identify NLG projects that are both publishable on standard benchmarks and meaningful for real-world, high-stakes applications?
4. Which research directions are most likely to advance NLG in the coming years?
5. How can we design evaluation methods for NLG that scale beyond small manual annotations while still capturing factual accuracy and reducing reliance on surface-level metrics?

4 Topics for Roundtable Discussions with Peers

1. The Role of Input Representation and Formatting in Generation Quality
2. Generalizing Factuality Evaluation: Bridging Closed Benchmarks and Real-World Complex Domains (sports, finance, or healthcare)
3. Scaling Evaluation: Designing Protocols that remain reliable while Reducing Cost and Annotation Effort
4. Navigating Career Paths from Academia: Opportunities and Challenges

Acknowledgements

We thank the anonymous reviewers for their constructive and detailed feedback, which helped clarify and improve this paper. Due to space constraints, some of their questions could not be addressed in the main paper; I will discuss them during the workshop and have included further clarifications and analyses in the appendices.

References

Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics* 11:227–249. https://aclanthology.org/2023.tacl-1.14.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, editors, *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, pages 131–137. https://doi.org/10.18653/v1/2020.inlg-1.19.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. https://arxiv.org/abs/2102.01672.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. https://arxiv.org/abs/2202.06935.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. https://arxiv.org/abs/2501.03200.

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.

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)*. Association for Computational Linguistics, Online, pages 1173–1186. https://doi.org/10.18653/v1/2020.emnlp-main.89.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. https://arxiv.org/abs/1910.10683.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery* 36:318 – 354. https://api.semanticscholar.org/CorpusID:231802211.

Barkavi Sundararajan, Somayajulu Sripada, and Ehud Reiter. 2022. Error analysis of ToTTo table-to-text neural NLG models. In Antoine Bosselut, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova,

and Laura Perez-Beltrachini, editors, *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pages 456–470. https://doi.org/10.18653/v1/2022.gem-1.43.

Barkavi Sundararajan, Yaji Sripada, and Ehud Reiter. 2024. Improving factual accuracy of neural table-to-text output by addressing input problems in ToTTo. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *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 7350–7376. https://doi.org/10.18653/v1/2024.naacl-long.408.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language* 80:101482. https://doi.org/https://doi.org/10.1016/j.csl.2023.101482.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .

Appendices

A Inter-Annotation Procedure for Participants

1. **Guidelines.** For the *ToTTo* and *NBA play-by-play summaries* annotations, I provided a detailed document defining the error categories with examples.

2. **Recruitment.** I initially recruited four annotators for the ToTTo task (Sundararajan et al., 2024); two did not proceed beyond the pilot stage after a quality check. Two trained annotators and I completed the final *ToTTo* annotations. For the NBA task, two qualified annotators started; one of them withdrew due to time constraints. The other annotator and I completed the annotations for *NBA game summaries*.

3. **Clarification.** Before the main task, I held short clarification sessions (meetings and emails) to ensure a shared understanding of the guidelines.

4. **Annotation.** Annotators completed their assigned samples independently and uploaded their annotation files via a Microsoft form.

5. **Compensation.** For the NBA task, the external annotator received a minimum payment of €35.

B Input Data Corrections and Output Text

We followed a systematic procedure to apply manual fixes identified in specific input problems in the ToTTo politics-domain dataset, as detailed in Sundararajan et al. (2024). A simple pseudocode is presented, where the Algorithm 1 takes tabular data and title (metadata) from ToTTo as input parameters to execute six functions to gather insights and make corrections to each of the input problems. The first two functions, DATASIZEMANAGEABLEFORLONGTABLE and IDENTIFYLEADERNAMEORDER do not correct the tabular input but provide insights on the input problems leading to factual errors in the outputs. The other four functions, CORRECTNONATOMICCELLS, UPDATEHEADERS, ADDRESSMISSINGVALUES, and REPLACESYMBOLS correct the input problems as presented in the paper.

Function summaries.

DATASIZEMANAGEABLEFORLONGTABLE(tabularData)
checks size thresholds (default: $\leq$ 20 rows, $\leq$ 10 columns); returns a boolean.

IDENTIFYLEADERNAMEORDER(tabularData, title)
extracts the leader name from the title and scans rows to record the name sequence; returns $(\mathit{leaderNameFromTitle}, \mathit{recordedData})$.

CORRECTNONATOMICCELLS(tabularData)
splits multi-valued cells into atomic columns; returns whether any corrections were made.

UPDATEHEADERS(tabularData)
fixes nested/misaligned headers; returns whether any corrections were made.

ADDRESSMISSINGVALUES(tabularData)
fills missing cells from local context; returns whether any corrections were made.

REPLACESYMBOLS(tabularData)
normalises domain-specific symbols and party abbreviations; returns whether any corrections were made.

Algorithm 1 Manual Correction Procedure for ToTTo Tabular Input

function MAINCORRECTIONPROCEDURE($tabularData, title$)
  if not DATASIZEMANAGEABLEFORLONGTABLE($tabularData$) **then**
    return "Please simplify the tabular data with fewer records", null
  end if
  $leaderName, recordedData \leftarrow$ IDENTIFYLEADERNAMEORDER($tabularData, title$)
  $correctionsMade \leftarrow$ **false**
  $correctionsMade \leftarrow$ CORRECTNONATOMICCELLS($tabularData$) **or** $correctionsMade$
  $correctionsMade \leftarrow$ UPDATEHEADERS($tabularData$) **or** $correctionsMade$
  $correctionsMade \leftarrow$ ADDRESSMISSINGVALUES($tabularData$) **or** $correctionsMade$
  $correctionsMade \leftarrow$ REPLACESYMBOLS($tabularData$) **or** $correctionsMade$
  if $correctionsMade$ **then**
    return $tabularData$, "Leader Data: " + $recordedData$ ▷ "Returns corrected input"
  else
    return $tabularData$, "Leader Data: " + $recordedData$ ▷ "Returns original input, if no issues"
  end if
end function

Algorithm 2 Verify the number of rows and columns in Longer Tabular Input

function DATASIZEMANAGEABLEFORLONGTABLE($tabularData$)
  $maxRows \leftarrow 20$ ▷ Maximum allowable number of rows for the chosen models
  $maxCols \leftarrow 10$ ▷ Maximum allowable number of columns for the chosen models
  return (length(*tabularData*) $\leq maxRows$) **and** (length(*tabularData*[0]) $\leq maxCols$) ▷ Both conditions must be true for returning true.
end function

Algorithm 3 Identify Leader Name Order in Title and Rows

function IDENTIFYLEADERNAMEORDER(*tabularData*, *title*)
  $leaderNameFromTitle \leftarrow$ ExtractLeaderNameFromTitle(*title*)
  $leaderNamesFromRows \leftarrow [\,]$
  for each row **in** *tabularData* **do**
    for each cell **in** current row **do**
      if leader_name is found in cell **then**
        Add leader_name to $leaderNamesFromRows$ ▷ Preliminary step: Record leader's names from rows
      end if
    end for
  end for
  if leaderNameFromTitle is not None **then** ▷ **Scenario 1:** Leader's name found in title
    $recordedData \leftarrow [\,]$ ▷ Initialize empty list to store sequential order of leader names
    Add $leaderNameFromTitle$ to $recordedData$
    for each leader_name **in** *leaderNamesFromRows* **do**
      Add leader_name to $recordedData$
    end for
    return $leaderNameFromTitle, recordedData$ ▷ for example, this returns "Leader b", ["Leader b", "Leader a", "Leader b", "Leader c"]
  else ▷ **Scenario 2:** Leader name not found in title
    if length of $leaderNamesFromRows > 0$ **then** ▷ when leader's name is found in at least one or more rows
      $recordedData \leftarrow$ "Leader name not in title"
      for each leader_name **in** *leaderNamesFromRows* **do**
        Add leader_name to $recordedData$
      end for
      return $leaderNameFromTitle, recordedData$ ▷ for example, this returns None, ["Leader name not in title", "Leader a", "Leader b", "Leader c"]
    else ▷ **Scenario 3:** Leader name not found in title or rows
      return $leaderNameFromTitle$, "Leader name not found"
    end if
  end if
end function

Algorithm 4 Correct Non-Atomic Cells

function CORRECTNONATOMICCELLS($tabularData$)
  $correctionsMade \leftarrow$ **false**
  for all $cell$ **in** $tabularData$ **do**
    if CellIsNonAtomic($cell$) **then**
      CorrectCell($cell$) ▷ Separate multiple atomic values into individual columns
      $correctionsMade \leftarrow$ **true**
    end if
  end for
  return $correctionsMade$
end function

Algorithm 5 Update Column Headers to Atomic cells

```
function UPDATEHEADERS(tabularData)
    correctionsMade ← false
    for all column in tabularData do
        if HeaderIsIncorrect(column) then
            UpdateHeader(column)                    ▷ Fix nested column and row headers
            correctionsMade ← true
        end if
    end for
    return correctionsMade
end function
```

Algorithm 6 Address Missing cell Values

```
function ADDRESSMISSINGVALUES(tabularData)
    correctionsMade ← false
    for all row in tabularData do
        if RowHasMissingValues(row) then
            FillCellValues(row)                     ▷ Pass the right missing cells from the table
            correctionsMade ← true
        end if
    end for
    return correctionsMade
end function
```

Algorithm 7 Replace Politics Domain-Specific Symbols and Abbreviate Party Names with correct semantics

```
function REPLACESYMBOLS(tabularData)
    correctionsMade ← false
    for all cell in tabularData do
        if CellContainsDomainSpecificSymbols(cell) then
            SubstituteSymbolWithEquivalent(cell)    ▷ Pass the right semantic replacing symbols in the input
        else if CellContainsPartyAbbreviations(cell) then
            AbbreviatePartyNames(cell)              ▷ Replace party name abbreviations
            correctionsMade ← true
        end if
    end for
    return correctionsMade
end function
```

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 11–14

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.

Nalin Kumar

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

nkumar@ufal.mff.cuni.cz
https://ufal.mff.cuni.cz/nalin-kumar

1 Research interests

General-purpose Large Language Models (LLMs) have demonstrated exceptional performance in diverse NLP tasks by leveraging massive datasets and computational resources, followed by fine-tuning or prompt-tuning for specific tasks Team et al. (2023); Dubey et al. (2024); Team et al. (2024). While many leading models are closed-source and heavily biased toward high-resource languages, particularly English, open-source alternatives have emerged with comparable performance. However, low-resource languages continue to face challenges due to limited pretraining data, high costs of creating linguistically rich corpora, and inefficiencies in tokenization and vocabulary. My main goal is to explore and find **efficient approaches** to **training LLMs** for **low-resource languages**. My previous research includes work on data-to-text generation, abstractive summarization, and linguistic entrainment in dialogue generation.

1.1 Building LLMs for Low-Resource Languages

My current research focuses on developing novel methods for training language models in low-resource language settings. I am particularly interested in designing generalizable approaches for training LLMs with limited data. One of my current works explores a modular training strategy that adapts multilingual models into monolingual ones.

In this work, we hypothesize that full model tuning is often unnecessary and propose a more modular approach. Specifically, our method involves first training a language-specific tokenizer and creating corresponding embeddings, followed by tuning only the non-embedding parameters. We conduct a comprehensive analysis across multiple scenarios, including multilingual-to-monolingual transfer and adaptation from high-resource to low-resource monolingual models. When applied to multilingual models, our method significantly reduces the number of tunable parameters and overall training time. Evaluation on natural language understanding (NLU) tasks shows that our approach improves performance when adapting mBERT to low-resource languages. Specifically, we use POS-tagging Nivre et al. (2020), NER Rahimi et al. (2019), and mask-filling task to evaluate our approach. We notice an increase of 1.5x in the mask-filling scores for our modular approach over full-tuning of the model, in the case of Scottish Gaelic language. The scores stay almost stagnant on the other two evaluation metrics. Additionally, we also experiment with replacing the embeddings with custom FastText embeddings, however, the results stay similar to the default configuration with mBERT embeddings. We perform similar experiments with Irish as well, which, unlike Scottish Gaelic, is present in the pre-training languages of mBERT. Consequently, using a randomly initialised transformer, we plan to investigate if the inclusion of a language in pretraining dataset helps with better initialisation of non-embedding representations for our modular training. We plan to further validate our method on decoder-based as well. We choose paraphrasing and summarization tasks to evaluate the models on their language generation capability. Since there is no gold data available, we plan to create silver test data using the NMT system and the available monolingual corpora. Specifically, for each data instance in the monolingual corpus, we will create its corresponding synthetic input using NMT systems and state-of-the-art English LLMs, depending on the downstream task.

1.2 Parameter-efficient training with Artificial Language Initialisation

In parallel, I am also working on parameter-efficient training methods through artificial language-based pretraining strategies. Prior studies Papadimitriou and Jurafsky (2020); Chiang and yi Lee (2022) demonstrate that models pretrained on non-linguistic data can achieve performance comparable to those trained on English sentences. We adapt the best performing approach followed by a parameter-efficient *pretraining* for language acquisition from limited data.

Our method initializes the model using token embeddings trained with a shallow model, followed by tuning only the non-embedding parameters on non-linguistic data to introduce structural biases. To initialize the model with structural biases, we experiment with pretraining on two artificial languages proposed by Papadimitriou and

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 15–17

Jurafsky (2023): the nested parentheses language (NEST) and the crossing parentheses language (CROSS). Subsequently, the model is frozen and further pretrained on the 10M-token BabyLM corpus using LoRA adapters. Experiments on small-scale dataset show that this approach leads to faster convergence while maintaining performance comparable to classic full-model pretraining.

1.3 Future Research

For my future research, I plan to enhance these approaches with linguistically informed curriculum training. A slightly more ambitious goal of my research plan is to develop multilingual language generation models by swapping the embeddings using existing embedding alignment approaches. Specifically, the non-embedding model parameters are treated as a universal language representation, while injecting language-specific information using just the embeddings. Additionally, I plan to experiment with Direct Preference Optimization (DPO) Rafailov et al. (2024) and delexicalized pretraining to further improve training efficiency in low-resource scenarios.

2 Questions for Senior Researchers

- How to land an internship and how to prepare for the interviews
- How should one usually decide whether a research idea should be pursued further, or when it's better to move on?
- Where do you think NLG research is heading in the next few years, and what areas seem most promising, especially in the era of LLMs?

3 Suggested topics for discussion

- Multilingual Text Generation: How important is the role of tokenization and vocabulary extension? How can we efficiently adapt the large vocabulary from a multilingual model to a language-specific monolingual model? How do we address catastrophic forgetting while finetuning a multilingual model on a lower-resource language?
- Data-efficient training of NLG systems: How reliable is training using synthetic data? How is the performance comparable to the gold data? How do we address problems for low-resource languages that are linguistically distant from any higher resource language? How can we evaluate NLG models for languages with minimal resources?

3.1 Acknowledgments

The study was supported by the Charles University, project GA UK No. 302425.

References

Cheng-Han Chiang and Hung yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. https://arxiv.org/abs/2109.03537.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* .

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. 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 4034–4043. https://aclanthology.org/2020.lrec-1.497.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 6829–6839. https://doi.org/10.18653/v1/2020.emnlp-main.554.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 8402–8413. https://doi.org/10.18653/v1/2023.findings-emnlp.563.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 151–164. https://www.aclweb.org/anthology/P19-1015.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* .

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* .

Biographical sketch

Nalin Kumar is a first-year PhD student at the Institute of Formal and Applied Linguistics (UFAL), Charles University in Prague. He is working on exploring efficient methods for Natural Language Generation (NLG) systems for his dissertation. He has published a couple of papers in NLG, one of which was presented at NAACL-Findings 2024. His other works include topics like Neural Machine Translation and Sentiment Analysis. He has also worked in text-to-data and data-to-text generation in low-resource languages for his Master's thesis at UFAL.

Author Index