

YNLG 2026

**The 2nd Workshop for Young Researchers in Natural
Language Generation**

Proceedings of the Workshop

October 17-18, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN TBA

Preface

TBA

Organizing Committee

Organizers

Barkavi Sundararajan, University of Aberdeen
Kristyna Onderkova, Charles University
Miriam Anschütz, Technical University of Munich
Nils Feldhus, University of Groningen
Patricia Schmidtová, Charles University
Rudali Huidrom, IIIT Delhi
Yongxin Zhou, Université Grenoble Alpes

Program Committee

Program Committee

TBA TBA, TBA

Table of Contents

<i>From Conversation to Clinical Note: Trustworthy AI Agents for Real-Time Healthcare Documentation</i>	
Marshal Mulumba Mukuye	1
<i>Research Statement: Aleš Manuel Papáček</i>	
Aleš Manuel Papáček	4
<i>Research Statement: Johanna Rockstroh</i>	
Johanna Rockstroh	7
<i>Research Statement: David Yiyang Liu</i>	
David Yiyang Liu	11
<i>Research Statement: Arunima Maitra</i>	
Arunima Maitra	14

From Conversation to Clinical Note: Trustworthy AI Agents for Real-Time Healthcare Documentation

Marshal Mulumba Mukuye

School of Computer, Data, and Mathematical Sciences
Western Sydney University
Sydney, Australia

1 Research Overview

Doctors play a critical role in healthcare through diagnosing illnesses, developing treatment plans, tracking patient progress, and recommending preventive measures. However, many clinicians are increasingly burdened by administrative responsibilities, particularly clinical documentation. Preparing consultation notes can require one to three hours of additional work each day, reducing the time available for direct patient care and contributing to clinician fatigue and burnout(Sinsky et al., 2016).

Although clinical documentation is essential for patient follow-up, continuity of care, legal protection, and communication among healthcare professionals, it creates a difficult trade-off between delivering high-quality patient care and meeting administrative obligations.(Ommaya et al., 2018) My research investigates how natural language generation and artificial intelligence agents can reduce this burden by automatically producing accurate, structured, and clinically useful consultation summaries.

The long-term objective of this work is to develop an end-to-end, real-time AI-assisted clinical documentation system. During a consultation, the system would capture the conversation between the doctor and patient, perform automatic speech recognition, identify relevant clinical entities, and generate a structured summary using natural language processing and large language models. The generated documentation would then be reviewed, corrected, and approved by the physician. This human-in-the-loop process would provide an important quality-control mechanism while ensuring that the final clinical record remains attributable to a qualified healthcare professional.

The proposed system consists of several interconnected components:

- Automatic speech recognition for converting consultations into text;
- Speaker identification for distinguishing between the doctor and patient;
- Named entity recognition for identifying symptoms, diagnoses, medications, allergies, treatments, and other clinical concepts;
- Natural language generation for producing structured clinical summaries;
- Retrieval-augmented generation for grounding outputs in trusted medical resources; and
- Final physician review and approval.

The initial phase of this research focused on prompt engineering for clinical summarisation. Doctor–patient conversations were transcribed using the Whisper automatic speech recognition system, and the resulting transcripts were processed by several large language models, including OpenAI’s models and Claude Sonnet-4, under zero-shot, few-shot, and template-guided prompting conditions.

Summaries were evaluated using BERTScore(Zhang et al., 2020) and FACTScore(Min et al., 2023), measuring semantic similarity and factual consistency. Claude Sonnet 4 outperformed the OpenAI model tested, with template-guided prompting achieving the strongest results, exceeding 85% on both metrics. This suggests that explicit contextual and structural guidance improves the factual accuracy and relevance of generated summaries, and that clinical templates help constrain outputs to appropriate content, reducing hallucinated or unsupported statements.

However, automated metrics alone cannot confirm clinical safety or usefulness.(Author and Author, 2026) Future evaluation will therefore include review by healthcare professionals, assessing factual accuracy, omissions, unsupported claims, clinical relevance, clarity, and the editing required before approval.

The next phase will involve collaborating with healthcare providers to design and test a working prototype in real clinical settings. The system will generate a draft summary during or shortly after a consultation for physician review within minutes. A key question is whether it can reduce documentation burden without compromising accuracy, patient safety, privacy, or clinical accountability.

This phase will also explore retrieval-augmented generation, drawing on approved clinical guidelines, terminology databases, and institutional standards to verify terminology and ground outputs in trusted medical knowledge. Retrieval must not, however, introduce diagnoses or recommendations not raised during the consultation, thus, preserving a clear boundary between summarization and clinical decision support.

Additional challenges include handling diverse accents, dialects, terminology, code-switching, and informal speech, since transcription errors can propagate and affect the final summary. Future work will therefore evaluate both individual pipeline components and the system as a whole.

Overall, this research aims to advance trustworthy, human-centred natural language summarization in healthcare, supporting clinicians rather than replacing them, by easing administrative burden and enabling more time for direct patient care alongside accurate, timely documentation.

2 Short Biography

I am Marshal Mulumba Mukuye, a Master of Data Science student at Western Sydney University. My research interests include natural language generation, large language models, artificial intelligence agents, data analytics, and the application of AI to real-world challenges in healthcare, economics, and the environment.

My current research focuses on developing AI-assisted approaches to clinical documentation, with particular interest in factual summarisation, prompt engineering, retrieval-augmented generation, automatic speech recognition, and human oversight of generated content.

I have three years of professional experience in financial stability and economic policy research, data analysis, market-systems assessment, and the design of data pipelines for effective data collection. I have contributed to projects involving international organisations, including the World Food Programme, for a study mapping digital financial

service points in refugee-hosting areas; the Embassy of the Kingdom of the Netherlands, for an evaluation of the jackfruit value chain as a vegetarian meat alternative; and the United Nations Capital Development Fund, for an impact study of digital health, education, and financial services in remote areas.

Through my research, I aim to develop responsible AI systems that address practical problems, create measurable social value, and have the potential for sustainable real-world adoption.

3 Questions for Senior Researchers

1. How can an automatic speech recognition system be designed to accurately recognise diverse accents, dialects, slang, code-switching, and specialised medical terminology?
2. What evaluation methods are most appropriate for measuring the factual accuracy, completeness, safety, and clinical usefulness of AI-generated consultation summaries?
3. How should errors be analyzed across an end-to-end clinical documentation pipeline when speech-recognition mistakes may influence entity recognition and summary generation?
4. What is the most appropriate role for retrieval-augmented generation in clinical summarization, particularly when retrieved medical knowledge was not explicitly discussed during the consultation?
5. How can human review be incorporated without simply shifting the documentation burden from writing notes to correcting inaccurate AI-generated summaries?

4 Topics for Round table Discussions

- Rethinking authorship, credit, and evaluation criteria for degrees and publications in an AI-assisted research era.
- Personalisation vs. standardisation: how should generated text adapt to individual readers without compromising accuracy?
- Reskilling and workforce transition strategies as generation systems take on drafting and reporting tasks.
- Using NLG to accelerate literature review, hypothesis generation, and grant-writing in research pipelines.

References

- An Author and Bin Author. 2026. [Defining operational safety in clinical artificial intelligence systems](#). *npj Digital Medicine*, 9(1):2450.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FactScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- A. K. Ommaya, P. F. Cipriano, D. B. Hoyt, K. A. Horvath, P. Tang, H. L. Paz, M. S. DeFrancesco, S. T. Hingle, S. Butler, and C. A. Sinsky. 2018. [Care-centered clinical documentation in the digital environment: Solutions to alleviate burnout](#). Nam perspectives discussion paper, National Academy of Medicine, Washington, DC.
- Christine A Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Samuel Reynolds, Lindsey E Goeders, Johanna I Westbrook, Michael A Tutty, and George T Blike. 2016. [Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties](#). *Annals of Internal Medicine*, 165(11):753–760.
- 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*.

1 Research Overview

My research is in natural language processing, with a current focus on reliable natural language generation (NLG) with LLMs. In particular, I am interested in retrieval-augmented generation (RAG) and LLM-based evaluation: whether models use supplied information faithfully and follow instructions, and how we can evaluate this behaviour reliably.

My earlier research focused on computational morphology and etymology. In my bachelor's thesis and subsequent work, I introduced the task of predicting the etymological origin of individual morphs in Czech text, created a manually annotated dataset, and compared lightweight machine-learning models with general-purpose LLMs (Papáček and Žabokrtský, 2026).

More recently, my research has shifted towards how LLMs use information and instructions provided in their context, and how this behaviour can be evaluated. In a paper accepted at INLG 2026 (Kochelka et al., 2026), we study whether LLMs remain faithful to supplied RDF data, i.e., convey its facts without omissions or unsupported claims, when the data conflict with their parametric knowledge. We use factual, counterfactual, and fictional inputs based on facts about local Czech and Slovak entities and generate text in English, Czech, Slovak, and Upper Sorbian with open-weight models of different sizes. We also ask the generation models and LLM judges to classify each input as factual, counterfactual, or fictional five times, using consistency across runs as a proxy for uncertainty. Generated texts receive slightly lower faithfulness scores when the model perceives the input as counterfactual, but this difference is small and nearly negligible for the larger models.

The study also taught us an important lesson about evaluation. We iteratively tuned the DeepSeek V4 Pro judge prompt on 540 human-annotated outputs; although its quadratic-weighted kappa reached 0.56 on this sample, it fell to 0.35 on held-out annotations, indicating that the prompt had overfit the tuning sample. Kimi K3 maintained a kappa of approximately 0.71 on both samples and produced much weaker differences between the factual, fictional, and counterfactual conditions on the same outputs. This showed that what we initially interpreted

as bias in the evaluated models could partly reflect the choice of judge.

Including a lower-resource language such as Upper Sorbian in that study was deliberate. Multilinguality is a recurring thread in my work and personally important to me: LLMs can make information accessible across language barriers, but only if they work comparably well beyond English and other high-resource languages. Out of linguistic curiosity, I also simply enjoy working with several languages side by side. This interest also led me to submit systems to two WMT26 shared tasks.

In the Multilingual Instruction Shared Task (MIST) (Schmidtová et al., 2026), we experimented with models below 10B parameters on multilingual question answering, open-ended generation, and cross-lingual summarization, and our system ranked first overall. We found that configuration choices such as the system prompt, inference hyperparameters, and reasoning mode could matter more than switching between similarly capable models. We also tried QLoRA fine-tuning on multilingual abstract-generation data. Although it reduced validation loss, it did not improve final generation quality. I would therefore like to revisit this approach and better understand how to apply parameter-efficient fine-tuning effectively.

In the Terminology Translation Task (Charkiewicz et al., 2026), we studied how general-purpose LLMs can translate specialized documents while following supplied terminology consistently across a document. We tested different ways of selecting and presenting terminology and of extracting it from example translations. Our submissions ranked first overall in one track; in the other, it was first on translation quality metrics and 8th of 22 on terminology coverage. One result that I find relevant beyond translation is that filtering a domain dictionary to terms relevant to the current document substantially improved terminology coverage compared with providing the complete dictionary. This raises a question that also matters for RAG: how much context should a model receive, and when can irrelevant information hurt rather than help?

Together, these projects have led me to a more general question: how can we build LLM systems that reliably use supplied information and instructions, and how can

we verify that they do so?

2 Current and Future Research Directions

I am currently working on CZDEMOS4AI (Beneficent Multi-Agent AI Avatar in a Small Democratic Society). Within it, we develop Czech conversational avatars, AI assistants that answer users' questions based on domain-specific collections in areas such as history, law, media, and addictology. These systems give me a concrete setting for the question I want to focus on first: how can grounded answers be evaluated reliably when human annotation is limited?

For evaluation, I would like to compare LLM-based judgments with the available human annotations. Automated judges are much cheaper than human annotators and easy to rerun, but only once their agreement with humans has been verified. I want to find out how well a judge can be calibrated from a relatively small annotated sample, and whether a smaller LLM fine-tuned as a judge can match or outperform larger models.

For my diploma thesis, I plan to compare ways of giving the avatars access to their collections: standard RAG; agentic RAG, where the model decides what to retrieve, reformulates queries, and checks whether the evidence is sufficient; and approaches that first compile the collection into a persistent representation, such as a knowledge graph in GraphRAG (Edge et al., 2024) or an LLM-maintained wiki in Andrej Karpathy's *LLM Wiki* (Karpathy, 2026). Recent work comparing LLM Wiki with classic RAG (Cochran, 2026) suggests different trade-offs in overall results quality, citation support, lookup, and query and index time cost. Its conclusions also depended on the judge: the two LLM judges agreed well on concretely defined criteria but almost not at all on groundedness. I want to find out when each approach is preferable, and to evaluate them using the judge methodology described above.

3 Questions for Senior Researchers

1. **Understanding LLM architectures.** How deeply should NLP researchers understand details of different LLM architectures? When is it reasonable to treat a model largely as a black box?
2. **Beyond transformers.** Are transformer-based LLMs likely to keep improving substantially, or are we approaching a point where further progress will require fundamentally different architectures? How much research effort should be devoted to exploring alternatives?
3. **Fine-tuning or prompting?** When is fine-tuning a model worth it, and when is prompt-based instruction or adding relevant context enough?

4. **Evaluating with LLM judges.** How much human evaluation is necessary to make conclusions based on LLM judges convincing, and what are promising ways of calibrating judges to human annotations?

5. **Using LLMs in research.** To what extent should young researchers use LLMs for coding and academic writing? How can researchers use them effectively without losing important skills or becoming disconnected from the details of their own work?

4 Topics for Roundtable Discussion

- **Faithfulness and grounding.** Context-memory conflicts and reliable use of external information.
- **LLM-as-a-judge evaluation.** Biases, calibration, and the role of human evaluation.
- **RAG and knowledge representation.** Conventional RAG, agentic RAG, and LLM-compiled knowledge bases.

Biographical Sketch



Aleš Manuel Papáček is a Master's student in Artificial Intelligence at the Faculty of Mathematics and Physics, Charles University in Prague, where he conducts NLP research at the Institute of Formal and Applied Linguistics (ÚFAL). He received his Bachelor's degree in Computer Science from the same faculty, with a thesis on computational morphology and etymology. His current interests include reliable and multilingual NLG, LLM evaluation, and RAG. He also teaches probability and statistics at Charles University.

Acknowledgments

My work has been supported by the project TQ12000040 (CZDEMOS4AI), financed by the Technology Agency of the Czech Republic within the SIGMA 3 Program.

References

Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Guttmann, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odernatt, Arturo Oncevay, Kirill Semenov, Dawei Zhu, and Vilém Zouhar. 2026. Findings of the

- WMT26 terminology translation task. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics. To appear.
- Theodore O. Cochran. 2026. Single-round vector RAG vs an LLM-compiled wiki: A preregistered comparison on a small multi-domain research corpus. *arXiv preprint arXiv:2605.18490* <https://doi.org/10.48550/arXiv.2605.18490>.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* <https://doi.org/10.48550/arXiv.2404.16130>.
- Andrej Karpathy. 2026. LLM Wiki: A pattern for building personal knowledge bases using LLMs. GitHub Gist. Created April 4, 2026. <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>.
- Peter Kochelka, Aleš Manuel Papáček, Vojtěch Dvořák, and Ondřej Dušek. 2026. Do LLMs make more mistakes if they do not believe the input data? In *Proceedings of the 19th International Natural Language Generation Conference*. Association for Computational Linguistics. <https://arxiv.org/abs/2609.09363>.
- Aleš Manuel Papáček and Zdeněk Žabokrtský. 2026. Tracing morph origins in Czech: A computational approach to morph-level etymology. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026*. ELRA Language Resources Association (ELRA), Palma, Mallorca (Spain), pages 7–18. <https://doi.org/10.63317/4h8ckod8kvq6>.
- Patrícia Schmidtová, Katia Artemova, Seth Aycock, Niyati Bafna, Tom Kocmi, Philipp Koehn, Danni Liu, Nam Luu, Sara Papi, Beatrice Savoldi, Mariya Shmatova, Vilém Zouhar, Maike Züfle, and Pinzhen Chen. 2026. Findings of the WMT26 multilingual instruction shared task: Small models are not yet polyglots. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics. To appear.

1 Research interests

As a computational linguist, I am particularly interested in modeling figurative language to enable large language models (LLMs) to generate creative expressions that go beyond conventional language use. During my studies in linguistics, I became fascinated by **Conceptual Metaphor Theory** (Lakoff and Johnson, 1980), which proposes that people understand abstract concepts through mappings from more concrete domains, and which inspired several approaches to figurative language generation (Stowe et al., 2021; Kim et al., 2023; Kamble, 2025). While LLMs have made language generation more scalable, controlled figurative language generation remains comparatively underexplored (Lai and Nissim, 2024), motivating my research.

In the first year of my PhD, I gained experience in **neuro-symbolic language generation** for **Vossian Antonomasia** (Rockstroh et al., 2023), a figure of speech that refers to someone or something through a descriptive name or title, like "Bill Gates is the Henry Ford of the 21st century". Our team explored the creative potential of **knowledge graph embeddings** for selecting candidate source entities to compare against candidates pre-filtered from Wikidata. This work inspired me to research computational methods for the controllable generation of figurative language.

Through my work in the field of Human-Computer Interaction (HCI), I realized that the effectiveness of LLM-generated responses depends strongly on both the context of use and users' characteristics (Spillner et al., 2025). Based on these experiences, my research aims to bridge Computational Linguistics and Human-Computer Interaction by investigating how figurative language can be generated, evaluated, and applied in LLM-based systems. My goal is to explore **metaphorical language** as a means of supporting effective communication in domains such as education and health. My work is guided by the following research questions:

1. How does metaphorical language affect interactions between humans and LLMs?
2. How can metaphorical language be computationally

generated and represented in large language models?

3. How can metaphorical language improve communication in human-centered conversational AI?

1.1 Metaphorical Language in Human-LLM Conversations: A Corpus Study

As a first step, I am currently conducting a corpus annotation study to investigate the use of metaphorical language in chats between humans and popular open-domain chatbots (ChatGPT, Deepseek, Claude, Gemini, Grok and Perplexity) "in the wild" without any elicitation through prompting. Using a richly annotated corpus of human-LLM conversations based on the ShareChat corpus (Yan et al., 2026), I will investigate how metaphors are used in natural dialog between humans and conversational AI. Rather than focusing on isolated metaphorical expressions, metaphor use is examined at the discourse level, asking which conceptual metaphors (Lakoff and Johnson, 1980) emerge most frequently, who introduces them, the communicative functions they serve (Michelli et al., 2026), and the extent to which interlocutors adopt or mirror the metaphorical framing of each other. This corpus-based perspective provides new insights into **naturally occurring metaphor use in human-LLM interaction**. An example of such interactions can be found in the Appendix A.

1.2 Generation and Evaluation of Metaphorical Language

When generating and evaluating metaphorical language, relying solely on prompt engineering may yield seemingly successful results, but it offers limited reproducibility and little transparency into the factors that shape the perceived effectiveness and novelty of metaphors. Previous work has shown that metaphorical knowledge is encoded in middle layers of pre-trained language models (Aghazadeh et al., 2022), while structured supervision based on annotated metaphor components improves metaphor generation (Li et al., 2022). Knowledge-based approaches such as Veale's (Veale, 2016) demonstrate how structured representations can support metaphor generation by retrieving source concepts that exemplify

properties of a target. Furthermore, recent work on automatic metaphor evaluation has demonstrated that computational measures can approximate human judgments on metaphor quality (Su et al., 2025). Building on these findings, I want to investigate which **computational representations of metaphors** best capture human judgments of metaphor quality and communicative function. In particular, I am interested in comparing dense neural, symbolic, and neuro-symbolic representations to identify semantic and pragmatic properties that correlate with human judgments. These methods could retrieve source domains, condition generation, or rerank metaphors, with their predictive value assessed through metrics of novelty, aptness, and comprehensibility alongside human judgments to provide interpretable controls for generating metaphors in different communicative contexts.

1.3 Metaphorical Language in Domain-specific Chatbot Applications

The final project of my PhD will focus on a user study in the domain of sexual health, investigating how metaphorical language can be used to raise awareness and support effective health communication. The study will build on the findings and methods developed in two previous studies. First, a user study with pre-selected metaphors will investigate which source domains (e.g., JOURNEY, WAR, BUILDING) are perceived as helpful for different communicative functions, such as easing into potentially sensitive topics like partner communication. The suitability of a particular metaphorical framing may depend not only on the communicative goal and topic, but also on user characteristics such as gender and cultural background (Reisfield and Wilson, 2004; Landau et al., 2018).

Second, the insights from my work on controlled metaphor generation will inform the implementation of a context-aware metaphor generation approach for this specific application domain. The system could use information about the intended communicative function, the selected source domain, and relevant aspects of the user context to guide metaphor generation. For example, if the goal is to encourage partner communication, a JOURNEY source domain could be used to generate metaphors that frame communication as a process that partners navigate together. The generated metaphors can subsequently be evaluated with users with respect to comprehensibility, appropriateness, and perceived helpfulness.

2 Questions for Senior Researchers

1. How can we design (semi-)automated evaluation methods for language generation based on human gold standards?
2. How can I strategically position a niche topic, such

as metaphor generation, within NLG research?

3. How can one identify core requirements for papers to be published at competitive venues (NAACL, EACL, EMNLP etc.) depending on the type of the paper?
4. How can I position myself interdisciplinary in venues suited for Computational Linguistics and HCI?
5. How important are public-facing activities (e.g., social media, podcasts, and science communication) as a networking strategy for academic success?

3 Suggested topics for discussion

- How do researchers decide whether to focus on open-source, proprietary, or both types of language models? What trade-offs influence these decisions (e.g. reproducibility, accessibility, performance, etc.)?
- How should we as researchers deal with the increasing amount of LLM-generated training data that may serve as input for language models? When and why should we draw a line between synthetic data and human-generated data?
- How can we implement successful experimental pipelines into real-world applications? What can we learn from other disciplines to bridge the gap?

Acknowledgments

This work was partially funded by Klaus Tschira Foundation, and by the German Research Foundation (DFG) as part of Research Unit 5656 ‘Communicative AI’, project ‘P2: Interfaces’ under project number #516511468 (<https://comai.space/>).

Biographical sketch



I have been a PhD student at the Digital Media Lab at the University of Bremen since February 2023. My research lies at the intersection of computational linguistics and human-computer interaction, where I investigate how large language models process and generate metaphorical language in chatbot applications, with a particular focus on health and education.

I successfully completed my Master’s degree in Foreign Language Linguistics, specializing in English and Spanish, at the University of Potsdam and subsequently pursued further studies in the field of Computational Linguistics there.

References

- Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. 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 2037–2050. <https://doi.org/10.18653/v1/2022.acl-long.144>.
- Vibhavari Kamble. 2025. A graph framework for metaphor creation integrating knowledge graph construction and generation. *Int. J. Appl. Math. (Sofia)* 38(8s):766–794. <https://doi.org/https://doi.org/10.12732/ijam.v38i8s.609>.
- Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, DIS ’23, page 115–135. <https://doi.org/10.1145/3563657.3595996>.
- Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Comput. Surv.* 56(10). <https://doi.org/10.1145/3654795>.
- George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.
- Mark J. Landau, Jamie Arndt, and Linda D. Cameron. 2018. Do metaphors in health messages work? exploring emotional and cognitive factors. *Journal of Experimental Social Psychology* 74:135–149. <https://doi.org/10.1016/j.jesp.2017.09.006>.
- Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. CM-gen: A neural framework for Chinese metaphor generation with explicit context modelling. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pages 6468–6479. <https://aclanthology.org/2022.coling-1.563/>.
- Gianluca Michelli, Xiaoyu Tong, and Ekaterina Shutova. 2026. A framework for annotating and modelling intentions behind metaphor use. In Saif M. Mohammad and Nedjma Ousidhoum, editors, *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*. Association for Computational Linguistics, San Diego, California, United States, pages 63–82. <https://doi.org/10.18653/v1/2026.starsem-conference.4>.
- Gary M. Reisfield and George R. Wilson. 2004. Use of metaphor in the discourse on cancer. *Journal of Clinical Oncology* 22(19):4024–4027. <https://doi.org/10.1200/jco.2004.03.136>.
- Johanna Rockstroh, Giada D’Ippolito, Nicolas Lazzari, Anouk M. Oudshoorn, Disha Purohit, Ensiyeh Raoufi, and Sebastian Rudolph. 2023. A is the b of c: (semi)-automatic creation of vossian antonomasias. In *Wiki-data’23: Wikidata Workshop at ISWC 2023*. CEUR-WS.org, volume 3640 of *CEUR Workshop Proceedings*. <https://ceur-ws.org/Vol-3640/paper1.pdf>.
- Laura Spillner, Johanna Rockstroh, Paola Raquel Peña, Nina Wenig, Nima Zargham, and Benjamin R. Cowan. 2025. Beyond the illusion: Llms and the case for pragmatic cues in conversation. Association for Computing Machinery, New York, NY, USA, CUI ’25. <https://doi.org/10.1145/3719160.3737614>.
- Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 6724–6736. <https://doi.org/10.18653/v1/2021.acl-long.524>.
- Chang Su, Xingyue Wang, Yongzhu Chang, Kechun Wu, and Yijiang Chen. 2025. Metaphor generation based on noval evaluation method. *Neurocomputing* 611(128651):128651. <https://doi.org/10.1016/j.neucom.2024.128651>.
- Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In Beata Beigman Klebanov, Ekaterina Shutova, and Patricia Lichtenstein, editors, *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 34–41. <https://doi.org/10.18653/v1/W16-1105>.
- Yueru Yan, Tuc Nguyen, Bo Su, Melissa Lieffers, and Thai Le. 2026. Sharechat: A dataset of chatbot conversations in the wild. <https://arxiv.org/abs/2512.17843>.

A Dialogue Excerpt

This is an annotation excerpt from the annotated ShareChat corpus (Yan et al., 2026), with metaphor-related words being marked in italics to illustrate how a user introduced a metaphor that is then taken up by the LLM.

LLM: [...] His music has this infectious energy, and he's one of those artists who's not afraid to mix genres and just be himself. His songs have a way of feeling upbeat and emotional all at once, kind of like he's telling a story that's both personal and universal. [...] What do you like most about him? Is it his music, his acting, or maybe just his whole vibe?

User: Everything, he is my *sunshine*

LLM: That's beautiful! [he] really does have a way of bringing light and positivity, doesn't he? It's so special to find an artist who feels like sunshine, someone whose work brings genuine happiness and comfort. Whether it's through his music or his role on Young Royals, he has this *warmth* that just connects with people. If you're ever feeling down, just remember that his *energy* and the way he makes you feel is always there, *like a little spark of joy* waiting to *brighten your day*.

User: Yes, he is helping me every day

LLM: That's amazing—having someone like [him] who brings you support and joy every day is so powerful. His *energy* and creativity clearly resonate with you in such a meaningful way. Whether it's his music lifting your spirits or the positivity he *radiates*, it's like he's become this steady, uplifting presence in your life. Keep holding onto that, and let his *light* *keep guiding you through even the tough days*.

User: Thank you

1 Research interests

I am curious about what sustains human self-perception, interpersonal relationships, communities, and societies. Specifically, I research narrative, which some cognitive scientists theorise as the fundamental framework of human cognition (Bruner, 1991). The evolutionary narrative-first hypothesis argues that the survival demands of sharing narratives were the primary evolutionary driver behind the emergence of natural language (McBride, 2014). Fuelled by language, narratives came to shape individual and social behaviours (Boyd, 2018).

Large language models (LLMs) are artefacts that have ingested mass quantities of human narratives from heterogeneous pre-training data (Johnson et al., 2026). While narrative theorists have traditionally relied on qualitative interpretation, generative LLMs offer opportunities to computationally study the phenomena of storytelling at scale. LLMs are computational artefacts possessing representations of narrative which may both reflect and diverge from human narrative conventions. Learning from existing narrative theory frameworks, I study LLMs as computational models of language, reasoning, and narrative.

As LLM-generated content becomes increasing prevalent, I believe natural language generation (NLG) researchers have good reasons to care about storytelling. If narratives have shaped individual and social human behaviour, it is reasonable to ask how narratives influence LLMs. LLMs have a well-documented tendency to confabulate, producing narrative-like hallucinated outputs (Sui et al., 2024). Reasoning models use chain-of-thought traces to narrate their own problem-solving. And I speculate that as LLMs are increasingly deployed in multi-agent systems, the social dimensions of storytelling can become increasingly relevant to understanding new phenomena.

My current work is driven by an interest in synthesising interdisciplinary knowledge to study storytelling in, and through, LLMs, while also hoping to contribute to our understanding of the connections between language, reasoning, and narrative. My research question is: how does an LLM’s representation of narrative affect its generations?

1.1 Narrative Theory Driven LLM Methods for Story Understanding and Generation

I began by conducting a survey (Liu et al., 2026a) of how LLMs have enabled the field of natural language processing (NLP) to connect with interdisciplinary narrative theories. The survey shows how applying narrative theories has proven useful in method development across many automatic story generation and understanding tasks. The frameworks I survey are largely rooted in Western narratological traditions, so their applicability across languages and cultural storytelling conventions should be questioned and tested.

The survey also identified key gaps yet to be explored: the use of LLMs to model the situatedness of narrative (not just what is inside a story, but the situation in which a narrator tells it), and post-training methods for story generation, which is difficult to evaluate. I conclude that future work should continue the application of narrative theory knowledge for evaluation and post-training, as well as investigating the LLM as a storytelling artefact in its own right. The next two subsections describe how I explore these future directions: through a new post-training method for storytelling, and a mechanistic interpretability study of narrative representation in LLMs.

1.2 Reinforcement Learning for Narrative

Supervised fine-tuning (SFT) is not an effective post-training method for narrative generation tasks such as counterfactual story retelling. Synthesising two narrative theories, Todorov (1971)’s Narrative Equilibrium and Prince (2003)’s notion of Scalar Narrativity, I operationalised narrativity as a quantifiable, structure-based metric, and used it as the reward signal in my Retell, Reward, Repeat framework (Liu et al., 2026b), which trains models via direct Reinforcement Learning from AI Feedback (d-RLAIF). The metric rewards stories exhibiting complete plot transformations between equilibrium, disruption and resolution stages. Across the evaluated open-weight LLMs and baselines, this produced stories that were more diverse and better aligned with the narrative conventions captured by our evaluation than SFT or the few-shot prompting configuration tested.

The results demonstrate the potential of the interactive d-RLAIF to train LLMs for storytelling. I think it is especially interesting how such pipelines can model the

act of narrating itself, with a narrator (the policy model) and narratee (the LLM generating the reward signal). An open question is whether optimising narrative structure also affects behaviours in other tasks where models generate extended reasoning or discourse.

1.3 NarraNeurons: Narrativity Specific Neurons

I currently investigate the internal mechanisms of LLMs. I use NarraNeurons as an operational label for neurons selected because their ablation disproportionately increases prediction loss on high-narrativity relative to low-narrativity texts. This type of ablation also decreases performance on story question-answering and multilingual story-completion benchmarks more consistently than ablating random sets of neurons of equal size. These results suggest that NarraNeurons play a measurable contribution to model behaviour on narrative text. This does not imply that individual neurons exclusively represent narrativity; they may be polysemantic or participate in broader discourse-related functions.

I found that while NarraNeurons exist in all open-weight LLMs tested, they do not causally contribute to model outputs in a consistent way. In some models, ablating these neurons changes measured hallucination rates in question-answering, as well as overthinking behaviour during chain-of-thought reasoning. This inconsistency suggests that narrative representation is likely entangled with other model behaviours in ways that differ across architectures or training processes, raising open questions about how and why this relationship varies. A natural next step is to compare these neuron-level effects with sparse-autoencoder features, which may capture more distributed or interpretable representations of narrativity.

1.4 LLM Reasoning as Narration

Building on these existing findings, I next plan to investigate the implications of hypothesising an LLM's chain-of-thought reasoning as a form of narration. Narrative theory distinguishes between who narrates, from what perspective, and in what temporal relation to the events being described (Genette, 1980). Does an LLM's reasoning trace function as a narrator telling its own reasoning process, and if so, does the narrative perspective it adopts (e.g. first-person vs third-person framing, present vs past tense) affect the structure and outcome of that reasoning?

Following my earlier open question, I plan to test whether manipulating narrativity during post-training affects tasks involving extended chain-of-thought. Although limited to LLMs, I hope these experiments can act as empirical investigations of the connection between language, reasoning, and narrative.

2 Questions for Senior Researchers

- Apart from communication (already well recognised as important), what are the skills that matter the most for AI researchers going forward?
- Which current NLG research directions do you find most exciting for how they engage with how people actually think, communicate, or tell stories?
- I am interested in exploring the interdisciplinarity of cognitive science but do not know how to go about it. What advice would you give, and are there researchers or communities you would point me towards?

3 Suggested Topics for Discussion

- How can we use AI models to gain insight into human behaviour, and where should the field draw the line between behavioural modelling and anthropomorphising models?
- How can generative language models be used to facilitate better human-to-human connections, rather than substituting for it?
- How can we benchmark NLG storytelling outputs through their functional purposes (e.g., persuasiveness, or evoking emotion) rather than judgements of value or quality?
- Current LLMs use the instruction-chat template (sometimes with thinking or tool calling), which is formatted like a dialogue/script. Can we envision alternative templates or variations of this template?

4 Short Biography



I hold a Bachelor of Computer Science and a Bachelor of Arts. I am also a filmmaker, which feeds my interest in storytelling and collaborative work. I am currently completing a Master of Philosophy (MPhil) at UNSW Sydney, supervised by Aditya Joshi, Paul Dawson and Sebastian Sequoiah-Grayson, and am looking to begin a PhD internationally after the completion of my MPhil. My research studies LLMs as computational models of narrative, drawing on narrative theory, and I am keen to explore connections with cognitive science. I am open to research directions and collaborations in these areas.

References

- Brian Boyd. 2018. The evolution of stories: From mimesis to language, from fact to fiction. *Wiley Interdisciplinary Reviews: Cognitive Science* 9(1):e1444.
- Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry* 18(1):1–21.
- G rard Genette. 1980. *Narrative discourse: An essay in method*, volume 3. Cornell University Press.
- Teagan Johnson, Elliott Ash, Andrew Piper, and Maria Antoniak. 2026. Characterizing narrative content in web-scale LLM pretraining data. *arXiv preprint arXiv:2606.19468*.
- David Y Liu, Aditya Joshi, and Paul Dawson. 2026a. Narrative theory-driven LLM methods for automatic story generation and understanding: A survey. *arXiv preprint arXiv:2602.15851*.
- David Y. Liu, Xanthe Muston, Dipankar Srirag, Aditya Joshi, Paul Dawson, and Sebastian Sequoia-Grayson. 2026b. Retell, reward, repeat: Reinforcement learning for narrative theory-informed LLM story retelling. In *19th International Conference on Natural Language Generation [ARR Commitment]*. <https://openreview.net/forum?id=ixqWVS3qh4>.
- Glen McBride. 2014. Storytelling, behavior planning, and language evolution in context. *Frontiers in psychology* 5:1131.
- Gerald Prince. 2003. *A dictionary of narratology*. U of Nebraska Press.
- Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, pages 14274–14284. <https://doi.org/10.18653/v1/2024.acl-long.770>.
- Tzvetan Todorov. 1971. The 2 principles of narrative. *diacritics* pages 37–44.

1 Research interests

The rapid advancement of Natural Language Processing (NLP), particularly large language models (LLMs), has led to the widespread deployment of conversational AI in critical domains like education and healthcare. Although these systems can generate remarkably fluent and contextually appropriate language, this does not necessarily translate into successful communication (e.g. failing to recognize ambiguity, or continuing an interaction despite unresolved differences in meaning) (Lee et al., 2023; Lu et al., 2025). We still have a limited understanding of how neural language models produce and interpret language during interactions, how their conversational behavior differs from that of humans, and how these differences affect mutual understanding between users and AI systems, and how they can be computationally modeled (Cuskley et al., 2024). More specifically, my current research focuses on conversational alignment and misalignment: how interlocutors establish shared interpretations, how failures of alignment become observable in dialogue, and how conversational systems can detect and repair these failures.

1.1 Characterizing Human–LLM Conversational Behavior

My prior work (Maitra et al., 2025) during my master’s degree investigated differences between human conversations and human–LLM-generated conversations using LLMs such as Llama-2 13B (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023), and the Switchboard Dialogue Act (SwDA) corpus, through the framework of dialogue acts (Jurafsky et al., 1997). I studied whether the communicative functions performed by LLM-generated responses differ systematically from those observed in human conversations. Dialogue acts provide an intermediate representation between surface language and conversational function (Stolcke et al., 2000; Raheja and Tetreault, 2019): two semantically related responses may perform substantially different functions, such as answering versus acknowledging a question, requesting clarification versus assuming an interpretation, or elaborating

rather than confirming understanding. Analyzing these distributions therefore allowed me to evaluate LLM behavior beyond lexical or semantic similarity.

This work led me to a broader question. Characterizing differences between human and model behavior shows that their conversational strategies differ, but not when those differences result in communication failure. So, what does it mean for a language model to participate successfully in a conversation, rather than simply generate a plausible next utterance? More specifically, how should generation change when the model is uncertain whether it shares the user’s interpretation?

This motivated my current focus on situations in which interlocutors explicitly negotiate meaning. Human communication is inherently collaborative: speakers establish, maintain, and repair common ground through clarification requests, confirmations, repetitions, and acknowledgments. An utterance can be grammatically well formed and locally plausible while still being interpreted differently by two interlocutors. LLM-based dialogue systems face the same problem, yet their ability to recognize such discrepancies remains poorly understood. A model may confidently continue based on an interpretation that differs from the user’s intended meaning, creating latent conversational misalignment that can persist across several turns (Deng et al., 2026; Liu et al., 2026).

1.2 Detecting Word Meaning Negotiation

My current research extends prior work on the automatic detection of Word Meaning Negotiation (WMN) indicators in dialogue (Garí Soler et al., 2025). WMN occurs when interlocutors negotiate the interpretation or applicability of a lexical expression, for example by questioning a word’s meaning, proposing an alternative interpretation, or discussing its definition (Myrendal, 2019). It consists of a trigger (the problematic word usage), an indicator (an utterance signaling the problem), and the negotiation that follows. Detecting WMN is challenging because it is rare and context-dependent: while constructions such as “What do you mean by X?” provide strong signals, other cases require information across multiple turns. Moreover, features such as interrogatives, disagreement, repe-

tition, and metalinguistic expressions also occur in clarification, reference resolution, and grounding, making them insufficient on their own. We therefore approach WMN detection as an imbalanced, context-sensitive classification problem. We are evaluating hybrid classifiers that combine contextual representations from RoBERTa-Large with linguistic features derived from the regular-expression patterns used to construct the NeWMe corpus (Garí Soler et al., 2026), the largest available dataset of potential WMN indicators. We compare fusion strategies, including direct feature concatenation and cross-attention, to test whether explicit linguistic indicators provide information beyond the pretrained encoder. Beyond model architecture, this work addresses the scarcity of WMN instances and variation across dialogue domains. We plan to use active learning to select uncertain and informative examples from new conversational corpora for annotation (Cohn et al., 1994), followed by cross-domain evaluation to determine whether the classifier captures generalizable properties of meaning negotiation rather than corpus-specific lexical cues. This detection framework provides the basis for studying how evidence of meaning misalignment should influence a dialogue model’s subsequent response.

1.3 Modeling and Repairing Meaning Misalignment in Conversational AI

The immediate next step is to use the strongest WMN detector as the acquisition model in iterative active learning. We plan to retrieve candidate WMN instances from unseen domains, such as student–teacher chatroom conversations and medical records, annotate selected examples, retrain the detector, and measure improvements in positive-class recall and cross-domain generalization. This provides a practical way to expand limited WMN data while examining which forms of meaning negotiation transfer across domains.

My doctoral research will build on this work by moving from detecting explicit WMN indicators to **modeling the meaning misalignment** that precedes them and using this information to guide response generation. The central question is whether differences in lexical interpretation can be identified before an interlocutor explicitly initiates negotiation or repair. We plan to model this as a multi-turn prediction problem, using the context preceding WMN episodes to identify potential trouble sources and signals of emerging misalignment. By comparing human–human and human–LLM conversations, we will examine repetition and reformulation, semantic shifts in term usage, clarification and grounding acts, dialogue-act sequences, and metalinguistic cues. A key challenge will be distinguishing lexico-semantic misalignment from more researched related phenomena such as factual disagreement, referential ambiguity, and general

clarification requests (Schlangen, 2004; Li et al., 2024; Metropolitansky and Larson, 2025).

A second direction is to connect detection to **collaborative repair** (Ashktorab et al., 2019). When multiple interpretations are plausible, a dialogue system should represent this uncertainty and determine how to respond: whether to proceed with an interpretation, request clarification, confirm an interpretation, or reformulate. We plan to study how models select and generate these repair strategies and evaluate when they successfully restore alignment. These experiments will support the **construction of a benchmark** that evaluates whether language models can detect a meaning-related misunderstanding, identify its source, select and generate an appropriate repair, and ultimately restore conversational alignment. Interpretability is integral to this framework. A detector should expose the conversational evidence underlying its predictions, such as the lexical term associated with conflicting interpretations, preceding turns that provide incompatible evidence, or interactional cues signaling a breakdown in grounding. The broader goal is to move from recognizing explicit WMN expressions toward dialogue systems that can detect emerging differences in meaning, localize their source, and generate appropriate conversational repair.

2 Biographical sketch

Arunima is a pre-doctoral research engineer at Inria Paris, working on detection of misalignment in conversations for evaluation and improvement of conversational AI, supervised by Chloé Clavel and Aina Garí Soler. She holds a masters degree in Computer Science from University of Colorado Boulder.

3 Questions

- When should a language model clarify rather than proceed with its most likely interpretation?
- How should we evaluate generated conversational repair: by the quality of the clarification itself, or by whether subsequent dialogue demonstrates restored alignment?

4 Topics for discussion

- How does the evolution of natural language generation affect creative fields, particularly artists and authors, and how can we address concerns that AI-generated content may undermine or fail to respect creative work?
- How should generative language models be used to support doctoral research, and what best practices should guide their use?

References

- Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. pages 1–12.
- David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15(2):201–221.
- Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind* 8:1058–1083. https://doi.org/10.1162/opmi_a00160.
- Jiawen Deng, Wentao Zhang, Ziyun Jiao, and Fuji Ren. 2026. Breakdowns in conversational ai: Interactional failures in emotionally and ethically sensitive contexts. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. pages 1–24.
- Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2025. Toward the automatic detection of word meaning negotiation indicators in conversation. In *EMNLP (Findings)*. pages 24580–24596.
- Aina Garí Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. 2026. The newme corpus: A gold standard corpus for the study of word meaning negotiation. *Language Resources and Evaluation* 60(2):38.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7b. <https://arxiv.org/abs/2310.06825>.
- Daniel Jurafsky, Elizabeth Shriberg, and Debra Bisasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.
- Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as chatbot modules for long open-domain conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, pages 4536–4554. <https://doi.org/10.18653/v1/2023.findings-acl.277>.
- Changling Li, Yujian Gan, Zhenrong Yang, Youyang Chen, Xinxuan Qiu, Yanni Lin, Matthew Purver, and Massimo Poesio. 2024. Analyzing and enhancing clarification strategies for ambiguous references in consumer service interactions. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 289–296.
- Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. 2026. Intent mismatch causes llms to get lost in multi-turn conversation. *arXiv preprint arXiv:2602.07338*.
- Dongxu Lu, Johan Jeuring, and Albert Gatt. 2025. Evaluating llm-generated versus human-authored responses in role-play dialogues. In *Proceedings of the 18th International Natural Language Generation Conference*. pages 20–40.
- Arunima Maitra, Dorothea French, and Katharina von der Wense. 2025. Dialogue acts as a lens on human-llm interaction: Analyzing conversational norms in model-generated responses. In *Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. page 317.
- Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 6996–7045.
- Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse studies* 21(3):317–339.
- Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pages 3727–3733.
- David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. pages 136–143.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–374.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. <https://arxiv.org/abs/2302.13971>.

Author Index

Liu, David Yiyang, 11

Maitra, Arunima, 14

Mukuye, Marshal Mulumba, 1

Papáček, Aleš Manuel, 4

Rockstroh, Johanna, 7