Teaching and Critiquing Conceptualization and Operationalization in NLP

Vagrant Gautam
Heidelberg Institute for Theoretical Studies
Heidelberg, Germany
`vagrant.gautam@h-its.org`

Abstract

NLP researchers regularly invoke abstract concepts like "interpretability," "bias," "reasoning," and "stereotypes," without defining them. Each subfield has a shared understanding or conceptualization of what these terms mean and how we should treat them, and this shared understanding is the basis on which operational decisions are made: Datasets are built to evaluate these concepts, metrics are proposed to quantify them, and claims are made about systems. But what do they mean, what *should* they mean, and how should we measure them? I outline a seminar I created for students to explore these questions of conceptualization and operationalization, with an interdisciplinary reading list and an emphasis on discussion and critique.

1 Introduction

Natural language processing and computational linguistics are relatively young, with the main professional organization (ACL), flagship conference (ACL), and the CL journal only being founded in the 1960s and 70s. As the field matures, we have entered a period of critical reflection about practices and definitions. For example, since the question of ethical NLP was raised (Hovy and Spruit, 2016), we now have dedicated workshops (Hovy et al., 2017), policies,[1] tutorials (Benotti et al., 2025), and teaching (Friedrich and Zesch, 2021) for it.

More recently, NLP researchers have begun to interrogate how we define and measure abstract concepts such as bias and understanding, revealing that proposed methods for measuring or mitigating bias are "poorly matched to their motivation" (Blodgett et al., 2020), and that current benchmarking practices in natural language understanding "obscure the abilities that we want our benchmarks to measure" (Bowman and Dahl, 2021). Understanding these conceptual debates and using them to

move the field forward are a fundamental skill for training NLP scholars rather than NLP engineers (Prasad and Davis, 2024). This is why I developed a seminar on conceptualization and operationalization in NLP, with the following learning objectives:

- Read and critique papers (both interdisciplinary and more conventional NLP papers)

- Critically evaluate aspects of conceptualization (defining an abstract concept) and operationalization (creating empirical measures of it, such as datasets and metrics) in NLP

- Design NLP projects in ways that address critiques and push the discipline forward

2 Structure

The seminar was offered to Masters students and advanced Bachelors students in computational linguistics and computer science at Saarland University. After some introductory sessions, we covered new concepts every two weeks, for which everyone was expected to read and discuss four papers: A *critique/conceptual* paper about the concept, to use as a lens to critique three other recent *content* papers in NLP that use that concept (see Section 3 for details on the readings). Students presented and led discussions about each concept in pairs. To hone students' skills at critiquing research, I assigned papers that demonstrate the skill and gave targeted feedback on assignments, following principles of scaffolded learning (Wood et al., 1976).

Introductory sessions. I used the first three sessions to establish expectations and course logistics, to teach reading, conceptualization and operationalization, and to give a sample presentation/discussion. Students chose their preferred concepts and dates, which informed the schedule and content for the rest of the class. As reading papers is part of the hidden curriculum (Wilson, 2024), I felt it was important to teach it explicitly with:

---

[1] `https://ethics.aclweb.org/`

Keshav (2007); Eisner (2009); Carey et al. (2020); Fruehwald (2022). Then we discussed definitions of conceptualization (Subramonian et al., 2023), operationalization (Steidl and Werum, 2019), and how they impact research as well as society. Finally, I presented the concept of *names* and led a discussion about it as a sample for the synthesis and critique I expected from students.

Course requirements. I used *assignments* with discussion prompts to encourage reading and thinking about the papers before student-led *presentation/discussions* about each concept, during which I also graded *engagement* from other students. Students could also write an optional *final report* designing a novel NLP project or re-imagining one of the content papers we read in a way that addressed our critique. See Appendix B for details.

3 Readings

A full list of concepts, critique papers and content papers is shown in Appendix A. For future iterations of the course, content papers can be swapped out for more recent or relevant work (e.g., the hype around DeepSeek's release made DeepSeek-AI et al. (2025) an obvious late addition to the *reasoning* readings). Importantly, the critique/conceptual papers, which are often interdisciplinary, hold up regardless of current NLP trends.

For *interpretability*, we read perspectives from a machine learning researcher (Lipton, 2018) and a philosopher (Krishnan, 2020), that critique interpretability as being a slippery concept and even unnecessary. Krishnan's (2020) distinction between causal and justificatory explanations grounded our readings of the popular and more technical papers, Patchscopes (Ghandeharioun et al., 2024) and RAVEL (Huang et al., 2024).

As for *explainability*, we read Miller (2019), which summarizes findings from philosophy, cognitive science, and psychology that are relevant to human explanations and thus useful for explainability in NLP. We also read 'Attention is not Explanation' (Jain and Wallace, 2019) and a response to it titled 'Attention is not not Explanation' (Wiegreffe and Pinter, 2019), which is itself a critique. Beyond the scientific content of these papers, reading them is also an important part of understanding the history of the field and important debates. Another such example is Bender and Koller (2020) and Michael (2020), which has prompted rich discussion in other pedagogical settings as well (Guerzhoy, 2024).

4 Encouraging Discussion and Critique

Direction and feedback were centrally important for students to develop the skill of critique. In their assignments, students were asked to synthesize commonalities and differences across papers and point out what authors did especially well, or any choices they were skeptical about. Some students had trouble with this at the beginning, instead just summarizing what each individual paper did, but targeted feedback and examples helped. By the end of the term, some students were even bringing up points covered in other literature that we had not read in this course, and I could send them paper recommendations rather than constructive feedback in my responses to their assignments.

Assignments were due before the sessions where we discussed the concept, so discussions were also high quality. I emphasized that all of our perspectives are important to co-create knowledge in the classroom (McCrae, 2024), and I explicitly graded presenters on their classroom management and whether they gave everyone an opportunity to participate, since discussion-heavy seminars can be challenging for some students. A third of student evaluations explicitly mentioned this as a positive. Offering a variety of topics and types of papers also allowed students in this "mixed" undergraduate-graduate and interdisciplinary setting to participate, as others have observed (Fosler-Lussier, 2008).

The diversity of students in terms of sociodemographic factors like gender and country of origin were also particularly interesting for discussions of sociodemographic concepts such as *bias* (Blodgett et al., 2020), *stereotypes* (Blodgett et al., 2021), and *names* (Gautam et al., 2024c). Students also made several connections across this cluster of topics. Finally, students practised turning critique into action through final project proposals designed to address critique and push the discipline forward.

5 Conclusion

I have presented a seminar designed to teach critical thinking about abstract concepts in NLP, with attention to both its social and technical aspects. Through readings, assignments, presentations, discussions, and final reports, students gained an appreciation for how decisions about conceptualization and operationalization can impact all aspects of our research, as well as public perceptions of NLP technology and tools.

Limitations

The primary limitation of this course is that it does not scale in its current form; a small classroom is essential for equal participation and for quality feedback with just one instructor. Additionally, it is challenging to balance a cohesive reading schedule (e.g., putting related topics such as *names*, *stereotypes*, and *bias* close together) with optimizing for student-preferred presentation dates. Finally, this course works best when students engage intellectually with the content themselves, rather than outsourcing their thinking to LLMs (Guest et al., 2025). This cannot be ensured, but only emphasized by the instructor and encouraged by creating a classroom environment that acknowledges and embraces friction as part of the learning process.

Ethics Statement

I designed this class from a standpoint of seeing NLP technologies as fundamentally sociotechnical (Selbst et al., 2019; Dhole, 2023), value-laden (Birhane et al., 2022), and situated rather than objective (Haraway, 1988; Waseem et al., 2021). I wanted to share this with NLP and CL students who are rarely given the opportunity to engage with such perspectives from science and technology studies (STS) and human-computer interaction (HCI) research, but may nevertheless go on to have power as system designers in academia and industry (Scheuerman and Brubaker, 2024). My pedagogy is informed by critical, feminist and engaged pedagogies (hooks, 1994), which are important to create a learning environment that feels fundamentally safe to be present in and share, and to broaden epistemological engagement beyond traditional ways of knowing in computer science and STEM (Raji et al., 2021). This is particularly important when discussing concepts such as race, class, gender, and so on, where *power* is an important consideration that often goes unexamined.

Acknowledgements

I am grateful to my students for their openness and feedback, and for letting me learn with them.

References

Haozhe An and Rachel Rudinger. 2023. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vagrant Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLOS ONE*, 16(1):1–28.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Luciana Benotti, Fanny Ducel, Karën Fort, Guido Ivetta, Zhijing Jin, Min-Yen Kan, Seunghun J. Lee, Minzhi Li, Margot Mieskes, and Adriana Pagano. 2025. Navigating ethical challenges in NLP: Hands-on strategies for students and researchers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 7–8, Vienna, Austria. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Maureen A. Carey, Kevin L. Steiner, and William A. Petri, Jr. 2020. Ten simple rules for reading a scientific paper. *PLOS Computational Biology*, 16(7):1–6.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Kaustubh Dhole. 2023. Large language models as SocioTechnical systems. In *Proceedings of the Big Picture Workshop*, pages 66–79, Singapore. Association for Computational Linguistics.

Jason Eisner. 2009. How to read a technical paper.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *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)*, pages 1905–1925, Online. Association for Computational Linguistics.

Eric Fosler-Lussier. 2008. Strategies for teaching "mixed" computational linguistics classes. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 36–44, Columbus, Ohio. Association for Computational Linguistics.

Annemarie Friedrich and Torsten Zesch. 2021. A crash course on ethics for natural language processing. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 49–51, Online. Association for Computational Linguistics.

Josef Fruehwald. 2022. Reading a technical paper.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024a. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024b. WinoPron: Revisiting English Winogender schemas for consistency, coverage, and grammatical case. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 52–66, Miami. Association for Computational Linguistics.

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024c. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *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)*, pages 1926–1940, Online. Association for Computational Linguistics.

Michael Guerzhoy. 2024. Occam's razor and bender and koller's octopus. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 128–129, Bangkok, Thailand. Association for Computational Linguistics.

Olivia Guest, Marcela Suarez, and Iris van Rooij. 2025. Towards critical artificial intelligence literacies.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

bell hooks. 1994. *Teaching To Transgress*. Routledge.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L. Griffiths, Max Kleiman-Weiner, Jiawei Han, Asli Celikyilmaz, and Yulia Tsvetkov. 2025. Cognitive foundations for reasoning and their manifestation in llms. *Preprint*, arXiv:2511.16660.

S. Keshav. 2007. How to read a paper. *SIGCOMM Comput. Commun. Rev.*, 37(3):83–84.

Maya Krishnan. 2020. Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy and Technology*, 33(3):487–502.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024. Do emergent abilities exist in quantized large language models: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5174–5190, Torino, Italia. ELRA and ICCL.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.

John McCrae. 2024. Co-creational teaching of natural language processing. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 33–42, Bangkok, Thailand. Association for Computational Linguistics.

Julian Michael. 2020. To dissect an octopus: Making sense of the form/meaning debate.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *Preprint*, arXiv:1712.00547.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. 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 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 105–118, Bangkok, Thailand. Association for Computational Linguistics.

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 515–525, New York, NY, USA. Association for Computing Machinery.

Hayley Ross, Kathryn Davidson, and Najoung Kim. 2024. Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 131–153, Miami, Florida, USA. Association for Computational Linguistics.

Jacqueline Rowe, Mateusz Klimaszewski, Liane Guillou, Shannon Vallor, and Alexandra Birch. 2025. EuroGEST: Investigating gender stereotypes in multilingual language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32062–32084, Suzhou, China. Association for Computational Linguistics.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Preprint*, arXiv:2304.15004.

Morgan Klaus Scheuerman and Jed R. Brubaker. 2024. Products of positionality: How tech workers shape identity concepts in computer vision. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. 2023. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087, Toronto, Canada. Association for Computational Linguistics.

Christina R. Steidl and Regina Werum. 2019. If all you have is a hammer, everything looks like a nail: Operationalization matters. *Sociology Compass*, 13(8):e12727. E12727 SOCO-1495.R2.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in

multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada. Association for Computational Linguistics.

Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164, Singapore. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How large language models are transforming machine-paraphrase plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andreas Waldis, Joel Birrer, Anne Lauscher, and Iryna Gurevych. 2024. The Lou dataset - exploring the impact of gender-fair language in German text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10604–10624, Miami, Florida, USA. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp. *Preprint*, arXiv:2101.11974.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Shomir Wilson. 2024. Documenting the unwritten curriculum of student research. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 1–3, Bangkok, Thailand. Association for Computational Linguistics.

David Wood, Jerome S. Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

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, Mexico City, Mexico. Association for Computational Linguistics.

A Full List of Readings

A selection of critique/conceptual papers (and possible content papers in parentheses) is presented by concept below. Content papers may be swapped out in future iterations of the course for more recent or relevant work (or simply so that the instructor can read something new).

Stereotypes. Blodgett et al. (2021)
(Ungless et al., 2023; Mitchell et al., 2025; Rowe et al., 2025)

Interpretability. Lipton (2018); Krishnan (2020); Saphra and Wiegreffe (2024)
(Ghandeharioun et al., 2024; Huang et al., 2024)

Explainability. Miller et al. (2017); Miller (2019); Bastings and Filippova (2020)
(Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Jain et al., 2020)

Paraphrases. Bhagat and Hovy (2013)
(Wahle et al., 2022, 2023; Sharma et al., 2023)

Bias. Blodgett et al. (2020)
(Sap et al., 2020; Goldfarb-Tarrant et al., 2021; Sap et al., 2022; Parmar et al., 2023)

Race. Field et al. (2021)
(Sap et al., 2019; Deas et al., 2023; Bourgeade et al., 2023)

Emergent abilities. Schaeffer et al. (2023)
(Wei et al., 2022; Lu et al., 2024; Liu et al., 2024)

Gender. Larson (2017); Devinney et al. (2022)
(Vashishtha et al., 2023; Waldis et al., 2024; Gautam et al., 2024b)

Generalization. Hupkes et al. (2023)
(Weissweiler et al., 2023; Muennighoff et al., 2023; Ross et al., 2024)

Names. Gautam et al. (2024c)

(Asr et al., 2021; Sandoval et al., 2023; An and Rudinger, 2023)

Reasoning. Kargupta et al. (2025)

(Wu et al., 2024; Gautam et al., 2024a; DeepSeek-AI et al., 2025)

B Course Requirements and Instructions

For 7 credits, you are required to do a presentation and write a final report. You will be graded on: Engagement (30%), presenting a concept and leading a discussion (35%), and your final report (35%).

For 4 credits, you are only required to do a presentation. You will be graded on: Engagement (40%), and presenting a concept and leading a discussion (60%).

Assignments.

- Read the critique paper(s) first

- Then read (or do a quick scan, if you're short on time) the content papers

- Write 3-5 bullet points of synthesis/critique (you may also write more if you're feeling particularly inspired)

  – Connect ideas / definitions / methods across different papers
  – Critique what's missing or suspicious about the content papers using what you learned from the critique papers
  – Name some arguments that a paper makes and tell me whether you are convinced by them or not, and why

- For maximum points, I want to see that you engaged with all 4 of the papers beyond repeating their content

- **Do not use LLMs**: Please respect the time and effort I put it into reading your responses and giving you feedback, by putting your own thoughts down

Engagement.

- Mixture of assignments and in-class participation

- Minimum 33% each required to pass

- Assignments are to encourage you to do the readings and come to class prepared, because this is a discussion-heavy class

- In-class participation includes asking questions, answering questions and discussion prompts, etc.

Presentation + leading a discussion.

- Timing: 1 hour - 1 hour 15 mins

- Content should synthesize the readings

  – Look for commonalities and differences across papers (including from other concepts that we previously covered)
  – Discuss conceptualization and operationalization of your concept, and how they differ/coincide across the content papers
  – Critically question all assumptions that papers make
  – Try to answer the question "*Why?*" Why are we doing this work in the field? For what purpose is it useful? etc.

- Don't present the entire time

  – Your audience has also read the papers
  – Engage them meaningfully at **least** every 10-15 minutes (more is fine!)
  – I encourage the use of questions and discussion prompts for this
  – Ensure that everyone gets a chance to speak

- Meta stuff (also graded!)

  – Have clear / visually appealing slides (not too crowded / text-heavy)
  – Show up early to set up
  – Message me in advance to remind me if your team will need special cables or if you will be presenting from my laptop

Report.

- Pick a concept (does not have to be the one you presented on) and design a new research project that addresses a gap in how it is currently conceptualized / operationalized according to critiques (you can also re-design one of the content papers we read instead of making a new one, if you want)

- Your title should explicitly state the concept you picked

- 4 pages not including references, 11 or 12 point font, ACL ARR template

- Structure

  - Abstract
  - Introduction (motivate the gap in the literature that you want to fill, i.e., explain what the gap is in conceptualization/operationalization, and why it matters)
  - Experimental methodology (explain how, concretely, you will address the problem - If you will propose a new dataset, how will you design it? If you are proposing a new method, how will you evaluate it? If you're doing a new evaluation, what is the evaluation metric you are going to propose/use and why? If you propose all of the above I expect all parts to be well-motivated)
  - Related work (search for 10 or more papers on your concept that are closely related, for which you will do a quick scan and weave them together into a coherent narrative that contextualizes your work in the broader landscape of NLP research; make sure to address how your work relates to and differs from them)
  - Expected impact (explain how you expect your work to benefit the research community / advance science!)
  - Conclusion

- **No AI, please!** I want to see your raw thoughts and writing so you can learn how to structure your thoughts in written form and so I can give you feedback for later, and I definitely do not want to see references to papers that don't exist