Practical Tools from Domain Adaptation for Designing Inclusive, Equitable, and Robust Generative AI

Anthony Sicilia
Khoury College of Computer Sciences
Northeastern University
sicilia.a@northeastern.edu

Malihe Alikhani
Khoury College of Computer Sciences
Northeastern University
m.alikhani@northeastern.edu

Abstract

Generative language technologies have become integral to everyday communication, shaping social interactions and informing critical decision-making processes in areas such as recruitment, healthcare, and education. However, they often struggle to grasp the "long tail" of data distributions — concepts less frequently observed during training — which could have significant repercussions. These models may marginalize underrepresented groups by failing to comprehend preferred communication styles, such as code-switching, or perpetuating societal biases like gender bias. Sectors like healthcare, education, and law, requiring personalization and exhibiting nuanced linguistic features, are also particularly affected when pre-trained models misconstrue or overlook "long tail" data concepts. While methods like distillation of smaller language models, active learning, and other bias mitigation strategies can augment traditional training techniques, a careful statistical analysis is essential for their effective application. This tutorial offers a comprehensive examination of how to develop equitable, robust, and inclusive language technologies using statistical tools from Domain Adaptation (DA) that catalyze positive social change. We will delve into strategies for bias mitigation, explore how to measure bias, and examine open problems in creating culturally-grounded and inclusive language technologies. Accompanying code notebooks and packages will be provided.[1]

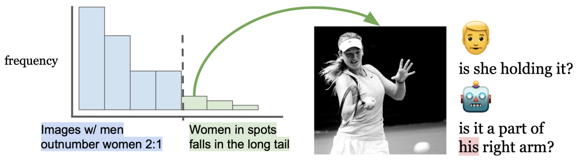

Figure 1: DA theory quantifies key properties of text data to inform us about model generalization; e.g., it can identify the long tail to promote equitable text generation for underrepresented groups.

1 Introduction

Large language models are increasingly deployed in critical areas of our daily life. Applications can improve health literacy (Ufuk, 2023), offer new avenues for improved education (Kasneci et al., 2023), and yield new legal technologies (Chalkidis et al., 2020). Meanwhile, as the complexity of these models increases, robust decision making, algorithm design, and evaluation become more and more important. It is vital that under-served demographics are not left behind in the wake of this technological wave – e.g., by supporting user-specific behaviors like code-switching (Harrington and Egede, 2023), low resource languages like American Sign Language (Inan et al., 2022), and equitable language use (Mayfield et al., 2019).

While we still have much to learn about new generative technologies (Rogers et al., 2020), what we do know can be alarming. For example, these models typically fail to learn infrequent data concepts in the long tail of text distributions (Kandpal et al., 2022). Indeed, this can lead to unfortunate, unintended outcomes such as social inequities (Bolukbasi et al., 2016), abysmal lexical diversity (Shekhar et al., 2019), or hard to resolve toxicity issues (Xu et al., 2021). All this is to say, without doubt, our use of machine learning as a tool has outpaced our understanding of this tool in many ways. For robust, responsible deployment of generative AI, we need a principled means of analysis. This tutorial aims to meet this demand, proposing domain adaptation (DA) theory as a mechanism to study the nuanced data issues that plague our models; e.g., the linguistic and societal biases induced by long tailed data. We cover statistical tools for:

1. *training* generative models with reinforcement learning, multi-agent techniques, distillation, traditional supervision, and more;

[1] https://github.com/anthonysicilia/AACL2023-DA4GenerativeAI

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Tutorial Abstracts), pages 14–22
November 1–4, 2023.

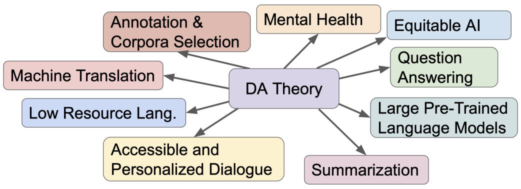

Figure 2: Overview of planned topics. Application of DA to text generation enables more than just obvious applications (e.g., model transfer). This tutorial focuses on these new emerging applications for generative AI.

2. *evaluating* the equity, human-likeness, inclusivity, and robustness of generative models;
3. and, *decision making* in small data regimes – e.g., model and dataset selection strategies.

Our accessible presentation of these tools can help to enable more robust deployment of generative AI.

While DA theory first appeared at *ACL venues over a decade ago (Blitzer et al., 2007), recently, more and more contemporary works have seen the benefit of carefully analyzing impacts of data-shift on their models. This is for good reason. DA theory allows us to answer complex questions like:

- *Will a pre-trained model generalize to my data?*
- *Can I improve generalization without much data?*
- *Is my corpus even large enough to measure bias and other language errors of a model?*

Despite its utility, the use of DA theory is not wide spread – a quick keyword search on `aclanthology` provided less than 10 papers at *ACL venues (excluding our own), which employ DA theory[2] or related techniques. This tutorial aims to bring awareness to emerging applications of DA theory to equitable, inclusive, and robust generation in an accessible way – connecting DA to more contemporary works whenever applicable.

2 Overview of Topics

We give a presentation plan next. For most topics, we highlight application areas or detailed questions we aim to address (see: ⋆ and italicized text), and also provide potential reading lists (see: ⧉).

1. **Inclusive Generation: Setup and Motivation**
   - Language Models and Data Sources
     ⋆ *Which end-users are left behind?*
     ⧉ Brown et al. (2020)
   - Example: Summarizing Medical Records
     ⋆ *Are domains with limited data impacted?*
     ⧉ Phan et al. (2023)
   - Example: Personalized Education
     ⋆ *Can we personalize generative models for individualized student experiences?*
     ⧉ Hu et al. (2008)
   - Example: Assistive Legal Technologies
     ⋆ *Can generative models be robust to specification (e.g., locality) in legal applications?*
     ⧉ Abdallah et al. (2023)
   - Example: Inclusive and Accessible Dialogue
     ⋆ *Can generative models support users with different preferences and capabilities?*
     ⧉ Sicilia et al. (2023); Inan et al. (2022)
2. **Domain Adaptation Theory: The Basics**
   - Learning Theory and Adaptation Bounds
     ⧉ Redko et al. (2020)
   - Classifier-based Statistical Distances
     ⧉ Ben-David et al. (2010); Sicilia et al. (2022a);
   - Measuring Model Data-Efficiency
     ⧉ Shalev-Shwartz and Ben-David (2014); Sicilia et al. (2021c)
   - Domain Adaptation for Generative Models
     ⧉ Sicilia and Alikhani (2022)
3. **Inclusive Text-Generation Algorithms**
   - Adversarial Training for Domain Alignment
     ⋆ *Application Areas: Unsupervised and Semi-supervised Summarization*
     ⧉ Ganin et al. (2016); Chen and Chen (2019)
   - Other Ways to Align: Semantics and Tokens
     ⋆ *Application Areas: Out-of-Domain Machine Translation and Low Resource Languages*
     ⧉ Štefánik et al. (2023); Phan et al. (2023)
   - Adapters and Adapter Soups
     ⋆ *Application Area: Adapting Language Models to New Domains without Training*
     ⧉ Chronopoulou et al. (2022, 2023)
   - Augmentation with Generative Models
     ⋆ *Applications: Semi-supervised Question-Answering, Accessible Dialogue, Counseling*
     ⧉ Yang et al. (2017); Parthasarathi et al. (2020); Shen et al. (2020); Inan et al. (2022)
   - Instance Weighting for Generative AI
     ⋆ *Applications: Out-of-Domain Machine Translation and Personalized Dialogue*
     ⧉ Wang et al. (2017); Welch et al. (2022)
   - Domain Adaptive MLM Objectives
     ⋆ *Applications: Mental Health Risk Prediction and other Healthcare Tasks*
     ⧉ Aragon et al. (2023); Lu et al. (2023)

[2] We distinguish between more common DA applications, and theoretical foundations; e.g., as in Redko et al. (2020).

4. **Computational Techniques (Activity)**
   - Confidence intervals and significance
     ⋆ *Is my test set large enough?*
     Ⓒ Shalev-Shwartz and Ben-David (2014)
   - Uncertainty and Confidence for Fairness
     ⋆ *Is my model fair to protected demographics? Do I even have enough data to determine this?*
     Ⓒ Ethayarajh (2020)
   - Transferring Models across Text-Genres
     ⋆ *How can I pick datasets when transferring models to small data regimes like medicine?*
     Ⓒ Blitzer et al. (2007); Atwell et al. (2022)
   - Supplementing Expertise with Bronze labels
     ⋆ *What's the best annotation protocol when (domain expert) gold labels are too expensive?*
     Ⓒ Hao and Paul (2019); Elsahar and Gallé (2019); He et al. (2021)
5. **Equitable Text-Generation**
   - Bias, Representational Harm, & Task Success
     Ⓒ Mayfield et al. (2019); Harrington and Egede (2023)
   - Defining Bias and Equity in Text-Generation
     Ⓒ Hendricks et al. (2018); Das and Balke (2022); Sicilia and Alikhani (2023)
   - Representation Learning and Bias Projection
     ⋆ *Applications: Mitigating Social Bias in Text Embedding and Masked Language Modeling*
     Ⓒ Vargas and Cotterell (2020); Yu et al. (2023); Kumar et al. (2023)
   - Data Augmentation and Interventions
     ⋆ *Applications: Toxicity Reduction in Masked Language Models and Equitable Distillation*
     Ⓒ Sun et al. (2019); Thakur et al. (2023)
   - Reinforcement Learning and Self-Play
     ⋆ *Applications: Morality, Toxicity, and Bias in Language Models; Bias in Dialogue Systems*
     Ⓒ Liu et al. (2022); Madanagopal and Caverlee (2023); Sicilia and Alikhani (2023)
6. **Future Work**: *TBA, Time Permitting*

3 Tutorial Type and Length

This tutorial is meant to be a **cutting-edge** tutorial and is meant to fill up a **3 hour time slot**.

Cutting Edge While DA theory has been well studied in ML theory communities, practical application for inclusivity, equity, and robustness of generative AI is an emerging area. Indeed, most of the reading-list has been published in *ACL venues across the last few years. While similar areas have been discussed in past tutorials (e.g., transfer learning and learning with limited data), the focus of this tutorial is on more rigorous theoretical aspects of DA and how these techniques can be applied in the, perhaps, unexpected area of equitable and inclusive generation. Our tutorial will also pay particular attention to large language models.

Timing We anticipate each of the 6 numbered top-level sections will take roughly 20 minutes, leaving extra time for questions and longer sections. Every 2 sections can be followed by a break.

4 Prerequisite Knowledge

Some familiarity with text-generation techniques and related tasks is recommended. The tutorial content will be accessible to Senior undergraduate, masters, and PhD students. In particular, **we assume no attendee will have experience with DA theory**, and plan to explain adaptation bounds and their distribution distances in an accessible way, giving preference to visualizations and high-level descriptions (over detailed equations). If desired, attendees can expound these topics themselves after the tutorial, using **take-home resources** provided during the talk or on the tutorial website (e.g., python packages, papers, surveys, etc.).[3]

5 Related Tutorials

No tutorial on DA theory for inclusive and equitable generation has been provided at an *ACL venue. With that said, recent tutorials have related motivation and complementary coverage.

Dyer et al. (2016); Church et al. (2022) and multiple other tutorials have previously considered deep neural networks for NLP. Deep networks have become a dominating trend and, as noted, their complexity poses issues for confident, responsible decision making as it pertains to training and deploying these models for generative applications. Our tutorial complements these existing tutorials, and pays careful attention to tools from DA theory specifically designed for large language models (Sicilia et al., 2022a). Our hope is to make application of these models more robust.

Chien (2019) present a tutorial on Deep Bayesian techniques, Ruder et al. (2019) present a tutorial on transfer learning, Yang et al. (2022) present a tutorial on learning with limited data, and

[3] https://github.com/anthonysicilia/AACL2023-DA4GenerativeAI

Fisch et al. (2022) present a tutorial on uncertainty estimation. These tutorials set the stage for our proposed tutorial, since DA theory provides rigorous solutions to many of the problems posed within these topics. As such, we do expect some topical overlap, but all of the techniques and solutions we present to attendees are likely to be new. Attendees that were/are interested in these previous tutorials will benefit from seeing how DA theory can be applied to solve their problems in a new way.

Tripodi and Pelillo (2016) present a tutorial on game theory, Belinkov et al. (2020) present a tutorial on interpretability, and Lucic et al. (2022) present a tutorial on reproducible ML. Each of these tutorials shares a common theme with our proposed tutorial: making NLP more robust through principled analyses. Similar to these tutorials, we will provide the tools for NLP practitioners applying ML to rigorously justify their decision making processes and algorithm designs.

Finally, Chang et al. (2019) present a tutorial on bias and fairness in NLP. Our tutorial complements this previous tutorial in topic, but presents a new perspective: the application of DA theory to this area with a focus on large generative models.

6 Instructors

Anthony Sicilia is a 5th year Ph.D student, specializing in applications of learning theory and domain adaptation theory to NLP problems such as inclusivity, equity, and robustness. He has experience in practical deployment of NLP systems, leading an Alexa Prize TaskBot team (focused on inclusivity and collaboration) to 3rd place overall in this international contest. He has published 4 papers on robust NLP at *ACL venues, which are present in the reading list: Atwell et al. (2022); Sicilia and Alikhani (2022); Sicilia et al. (2022b); Sicilia and Alikhani (2023). He also received a **best paper award** at UAI 2022 for his work on novel PAC-Bayesian DA theory for multiclass neural-networks (Sicilia et al., 2022a). His work spans application of DA theory to diverse areas such as: analysis of the impact of data-shift on parsers and sentiment classifiers, dialogue management and generation in non-cooperative multi-objective environments, causal analysis of the impact of model/dataset properties on discourse analysis, human-like dialogue management and generation, equitable dialogue management and generation, evaluation of both human-likeness and equity in dialogue, and quantification of linguistic and social biases in large language models. Previously, he also applied learning theory in vision, especially small-data medical applications with a primary focus on bias mitigation and robust model evaluation (Sicilia et al., 2021a,b,c; Zhao et al., 2022).

Malihe Alikhani is an expert in natural language processing (NLP) and machine learning. Alikhani's research interests center on using representations of communicative structure to improve ethical and practical machine learning models. One of the main focuses of her recent research has been on studying formal methods of machine learning for designing equitable and robust NLP tasks. This includes using tools from learning theory for efficient dialogue management, text generation, classification and measuring and mitigating biases in generation and classification tasks (Atwell et al., 2022; Sicilia and Alikhani, 2022; Sicilia et al., 2022b; Atwell et al., 2021; Sicilia and Alikhani, 2023; Sicilia et al., 2022a). Her work in these areas have received three best paper awards at UAI 2022, ACM UMAP 2022 and INLG 2021.

She has designed several task-oriented dialogue systems and conversational QA models (Khalid et al., 2020b,a; Sicilia et al., 2022b). Her work has explored data-driven modeling of inferential links in text and imagery (Alikhani and Stone, 2019), neural controllable description generation models for images (Alikhani et al., 2020b), datasets and models of coherent diagram interpretation (Alikhani and Stone, 2018; Hiippala et al., 2021) and interpretation of multimodal pointing actions in human-robot collaboration (Alikhani et al., 2020a). She has worked on distributional semantic approaches for modeling lexical aspect of verbs in English and six other languages (Kober et al., 2020). She has also been involved in various projects for studying the cognitive science of language use (Persaud et al., 2017) and formal language and automata, including probabilistic models of success runs in Markov independent trials (Alikhani et al., 2015). Alikhani has collected several corpora annotated by crowdworkers and expert linguists in the area of discourse, multimodality, dialogue, human–robot interaction and psycholinguistics (Alikhani and Stone, 2019; Hiippala et al., 2021; Alikhani and Stone, 2018; Alikhani et al., 2019a, 2020a). She has designed software for annotation, formal and ML models for studying communicative intents and the context of human-machine communication (Alikhani et al., 2019b; Khalid et al., 2020b).

References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *arXiv preprint arXiv:2304.06623*.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019a. CITE: A corpus of image-text discourse relations. 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 570–575.

Malihe Alikhani, Baber Khalid, Rahul Shome, Chaitanya Mitash, Kostas Bekris, and Matthew Stone. 2020a. That and there: Judging the intent of pointing actions with robotic arms. 34(06):10343–10351.

Malihe Alikhani, Bjørn Kjos-Hanssen, Amirarsalan Pakravan, and Babak Saadat. 2015. Pricing complexity options. *Algorithmic Finance*, 4(3-4):127–137.

Malihe Alikhani, Ethan Selfridge, Matthew Stone, and Michael Johnston. 2019b. Multimodal decisions for conversational ai. In *Submission*.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020b. CLUE: Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535.

Malihe Alikhani and Matthew Stone. 2018. Arrows are the verbs of diagrams. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3552–3563.

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a coherence relation: Evidence and implications. In *Second Workshop on Shortcomings in Vision and Language (SiVL)*.

Mario Aragon, Adrián Pastor López Monroy, Luis Gonzalez, David E Losada, and Manuel Montes. 2023. Disorbert: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1):151–175.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Francine Chen and Yan-Ying Chen. 2019. Adversarial domain adaptation using artificial titles for abstractive title generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2197–2203, Florence, Italy. Association for Computational Linguistics.

Jen-Tzung Chien. 2019. Deep Bayesian natural language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 25–30, Florence, Italy. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. 2023. Adaptersoup: Weight averaging to improve generalization of pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2009–2018.

Kenneth Church, Valia Kordoni, Gary Marcus, Ernest Davis, Yanjun Ma, and Zeyu Chen. 2022. A gentle introduction to deep nets and opportunities for the future. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dublin, Ireland. Association for Computational Linguistics.

Mayukh Das and Wolf Tilo Balke. 2022. Quantifying bias from decoding techniques in natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1311–1323, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chris Dyer, Yoav Goldberg, and Graham Neubig. 2016. Practical neural networks for NLP: From theory to code. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Austin, Texas. Association for Computational Linguistics.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. 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 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919, Online. Association for Computational Linguistics.

Adam Fisch, Robin Jia, and Tal Schuster. 2022. Uncertainty estimation for natural language processing. In *COLING*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Shudong Hao and Michael J. Paul. 2019. Analyzing Bayesian crosslingual transfer in topic models. 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 1551–1565, Minneapolis, Minnesota. Association for Computational Linguistics.

Christina N Harrington and Lisa Egede. 2023. Trust, comfort and relatability: Understanding black older adults' perceptions of chatbot design for health information seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2021. Foreseeing the benefits of incidental supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1782–1800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pages 771–787.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55(3):661–688.

Dawei Hu, Wei Chen, Qingtian Zeng, Tianyong Hao, Feng Min, and Liu Wenyin. 2008. Using a user-interactive qa system for personalized e-learning. *International Journal of Distance Education Technologies (IJDET)*, 6(3):1–22.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan, and Matthew Stone. 2020a. Discourse coherence, reference grounding and goal oriented dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Baber Khalid, Malihe Alikhani, and Matthew Stone. 2020b. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Thomas Kober, Stone Matthew Alikhani, Malihe, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Keming Lu, Peter Potash, Xihui Lin, Yuwen Sun, Zihan Qian, Zheng Yuan, Tristan Naumann, Tianxi Cai, and Junwei Lu. 2023. Prompt discriminative language models for domain adaptation. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 247–258.

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, and Robert Stojnic. 2022. Towards reproducible machine learning research in natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–11, Dublin, Ireland. Association for Computational Linguistics.

Karthic Madanagopal and James Caverlee. 2023. Reinforced sequence training based subjective bias correction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2585–2598, Dubrovnik, Croatia. Association for Computational Linguistics.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.

Prasanna Parthasarathi, Sharan Narang, and Arvind Neelakantan. 2020. On task-level dialogue composition of generative transformer model. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 41–47.

Kimele Persaud, Brian McMahan, Malihe Alikhani, Kevin Pei, Pernille Hemmer, and Matthew Stone. 2017. When is likely unlikely: Investigating the variability of vagueness. In *Proceedings of the Cognitive Science Society Conference*.

Long Phan, Tai Dang, Hieu Tran, Trieu Trinh, Vy Phan, Lam Chau, and Minh-Thang Luong. 2023. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3123–3134.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. 2020. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. 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 2578–2587, Minneapolis, Minnesota. Association for Computational Linguistics.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Anthony Sicilia and Malihe Alikhani. 2022. Leather: A framework for learning to generate human-like text in dialogue. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 30–53.

Anthony Sicilia and Malihe Alikhani. 2023. Learning to generate equitable text in dialogue from biased training data. In *Annual Meeting of the Association for Computational Linguistics 2023*.

Anthony Sicilia, Katherine Atwell, Malihe Alikhani, and Seong Jae Hwang. 2022a. Pac-bayesian domain adaptation bounds for multiclass learners. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Anthony Sicilia, Jennifer C Gates, and Malihe Alikhani. 2023. How old is gpt?: The humbel framework for evaluating language models using human demographic dat. *arXiv preprint arXiv:2305.14195*.

Anthony Sicilia, Tristan Maidment, Pat Healy, and Malihe Alikhani. 2022b. Modeling non-cooperative dialogue: Theoretical and empirical insights. *Transactions of the Association for Computational Linguistics*, 10:1084–1102.

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. 2021a. Domain adversarial neural networks for domain generalization: When it works and how to improve. *arXiv preprint arXiv:2102.03924*.

Anthony Sicilia, Xingchen Zhao, Davneet S Minhas, Erin E O'Connor, Howard J Aizenstein, William E Klunk, Dana L Tudorascu, and Seong Jae Hwang. 2021b. Multi-domain learning by meta-learning: Taking optimal steps in multi-domain loss landscapes by inner-loop learning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 650–654. IEEE.

Anthony Sicilia, Xingchen Zhao, Anastasia Sosnovskikh, and Seong Jae Hwang. 2021c. Pac bayesian performance guarantees for deep (stochastic) networks in medical imaging. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 560–570. Springer.

Michal Štefánik, Marek Kadlcik, and Petr Sojka. 2023. Soft alignment objectives for robust adaptation of language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8837–8853, Toronto, Canada. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Rocco Tripodi and Marcello Pelillo. 2016. Game theory and natural language: Origin, evolution and processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany. Association for Computational Linguistics.

Furkan Ufuk. 2023. The role and limitations of large language models such as chatgpt in clinical settings and medical journalism. *Radiology*, 307(3):e230276.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2902–2913, Online. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Diyi Yang, Ankur Parikh, and Colin Raffel. 2022. Learning with limited text data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 28–31, Dublin, Ireland. Association for Computational Linguistics.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada. Association for Computational Linguistics.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048.

Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. 2022. Test-time fourier style calibration for domain generalization. *arXiv preprint arXiv:2205.06427*.