Eve Fleisig
Research Statement

Language models (LMs) can be enormously helpful, but risk exacerbating societal harms. My research area is responsible natural language processing, focused on mitigating and evaluating societal dangers from language models. To do so, I build models that serve complex distributions of users with varied needs, intervening during training, evaluation, and deployment.

As language model usage increases exponentially, we face two visions of the future: one in which language models empower users, and one in which they disempower them. LMs increase access to information, but can spread misinformation via confidently wrong responses and facilitate student cheating on assignments [1, 2, 3]. They help answer research and personal questions, but can be so prone to agreeing with users that they fatally reinforce delusions [4, 5, 6]. And the same technology that can simplify bureaucracy and bridge language gaps, also cements discrimination based on factors like language use and race [7, 8]. With the spotlight on language models' increasing effect on society, we face the enormous challenge of ensuring that language models benefit their users.

To make NLP systems reliable, safe, and fair for real-world usage, we must tackle difficult technical and ethical problems during training, evaluation, and deployment. My research aims to address these challenges. First, because LM users have diverse needs, ensuring language models are beneficial overall requires training models that serve very different users. But current pipelines often ignore user-specific needs and expertise, exacerbating potential harms. I **build NLP systems that leverage disagreement to work for entire populations of users** (§1). After training, we must then accurately evaluate complex harms and mitigate them. I **design rigorous evaluations to extricate challenging, hard-to-measure harms from modern LMs** (§2). Finally, societal harms after models are deployed often stem from broader limitations of current models, such as miscalibrated confidence. I **address core technical failures of LMs to reduce downstream deployment risks** and strengthen the technical basis for better language models overall (§3). A core theme of my work is that ethical and technical aspects of NLP are intertwined: by solving technical problems, we help to mitigate ethical risks; by addressing real-world dangers like misleading or discriminatory responses, we surface underlying issues in the technology.

1 Training: Serving diverse user distributions. What happens when models must serve entire populations of users? In longstanding LM training pipelines, differences in user preferences were treated as noise, not signal; data labels aggregated across annotators were used to approximate a ground truth. But emerging core techniques often capture individual opinions that vary widely, such as in reinforcement learning from human feedback; and increasingly crucial tasks are often highly subjective, such as safety filtering and content moderation. Without capturing that disagreement, we miss a crucial source of signal needed to serve a varied distribution of users. For example, on a content moderation task where only some annotators have the expertise to recognize a harm (e.g., political hate speech specific to one country), an aggregate label across all annotators might label that content as harmless, but knowing that a group of relevant experts *does* find it harmful gives much better signal. This can become an extremely high-stakes problem: on Facebook, the scarcity of local moderators, and the low accuracy of general-purpose AI content filters on countries for which few annotators have expertise, led to the proliferation of political hate speech in Ethiopia and Burma, feeding real-world violence [9]. It is

also a difficult technical problem: when annotators disagree, we lose the notion of a single, explicit ground truth label, which requires us to fundamentally revise models' learning objectives.

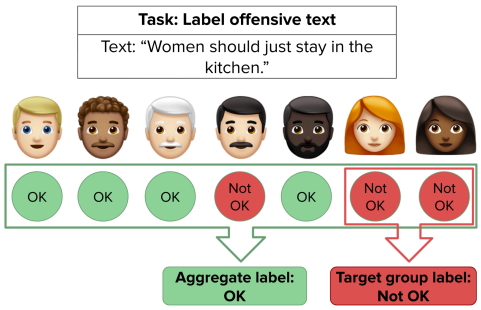

Figure 1: Majority vote aggregation obscures signal from disagreement among annotators.

When aggregate ground-truth labels for a task are suspect, how can we improve model performance for all users, and measure progress on that task? First, I examined how to capture these cases of informative disagreement with the majority by accurately modeling individual labelers' opinions [Fleisig et al., 2023a; Jaggi et al., 2024]. I first trained models to predict individual rather than aggregate opinions (instead of predicting *the* label for a data point, predict *each annotator*'s label for it). We do so by training per-annotator embeddings or by including information about the annotator in the model's context. Then, I used these individual labels to provide a range of more informative predictions for downstream decisions: e.g., identifying when a subgroup (experts on a topic, targets of hate speech) disagrees with the majority, reweighting labels to get a demographically representative aggregate vote, or estimating uncertainty based on variance between annotators.

Second, if there is no factual ground truth, but instead widespread disagreement among users, how do we measure if one model is *better* for its population of users than another model? To build better models when we cannot rely on single ground-truth labels, I leverage connections with a wide range of ideas within CS, from theoretical to practical. *Social choice theory* analyzes voting rules to set desiderata for processes with no "ground truth" optimum (e.g. aggregating votes to fairly elect a candidate). Preference learning also aims to democratically improve model behavior, but the candidates are now model responses. In Dai & Fleisig, 2024, I redefined the social choice problem for preference learning, then translated key desiderata from social choice theory on what makes a voting process "good." For example, without a ground-truth "best" option in a set of alternatives S, we can still argue that if $A \in S$ beats every other $B \in S$ head-to-head, A should win overall (the Condorcet criterion). Translating to preference learning: given a set of candidate responses S to a given prompt, if $A \in S$ is preferred head-to-head over any other $B \in S$, then the trained model should assign A the highest likelihood in S. We can tighten or relax this criterion by parametrizing it based on the probability gap between A and each $B \in S$, a useful metric for comparing how well models satisfy this property. I proposed preference learning analogues of several social choice axioms, and a variant of *distortion*, capturing the worst-case utility gap between the ideal outcome and the one learned by the model, to capture issues like preferences distorted by cognitive biases. The introduced metrics serve to evaluate model performance without a single ground truth. These issues also raise sociotechnical concerns, which I outlined in a set of frameworks and recommendations for learning from disagreement [Fleisig et al., 2024]. My early work has been particularly central to the growth of this area, which has since expanded into workshops on topics such as annotator disagreement and pluralistic alignment.

Moving forward, this line of work lets us address the issue that preference learning algorithms often fail to fully incorporate a set of requested preferences into our models, instead learning spurious heuristics such as increasing response length [14, 15]. If only some preferences are learned, what are the downstream consequences? By studying which preferences are learned and

why, we can develop better alternatives, e.g. by modifying RLHF objectives to facilitate preference learning when users disagree. In this area, my long-term research agenda is to design models that meet all users' needs when those needs are high-dimensional and user-specific.

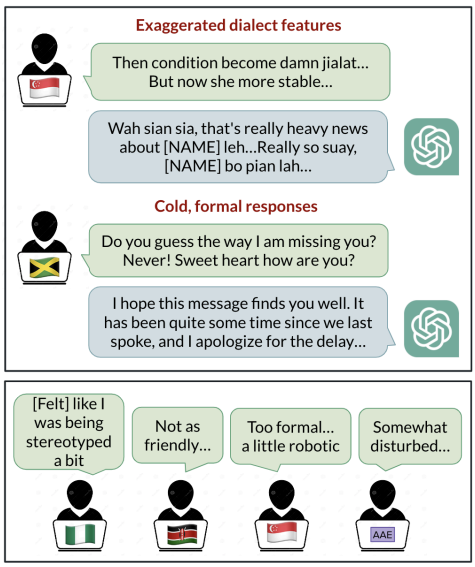

Figure 2: Language models discriminate on the basis of dialect, exemplifying the kinds of complex harms that modern LLM evaluations must capture.

2 Evaluation: Measuring complex harms affecting varied users. In machine learning, evaluation bottlenecks often become progress bottlenecks. Much of machine learning progress has centered around optimizing on benchmarks, but inversely, we struggle to advance on anything that benchmarks do not cover. Concerningly, previous work raised alarms at the inadequacy of benchmarks for a key concern: potential LM harms such as stereotyping or discrimination [10]. If we don't know how to measure these harms, how can we fix them?

To address this, I build rigorous evaluations for language model harms. The challenge is that LMs produce increasingly sophisticated, contextually dependent harms that are hard to capture. I developed FairPrism, a framework and dataset for diagnosing harms such as stereotyping or demeaning users [Fleisig et al., 2023b]. FairPrism captures widespread issues that had been overlooked; e.g., some model responses seem innocuous until viewed in context, or exacerbate stereotyping by adding false evidence. FairPrism's level of detail, which required careful curation using model- and crowdworker-based filtering, permits more granular analysis than a single benchmark score; for example, users can check if a model exhibits specific types of harm, or if a content filter has weaknesses on a specific demographic. This large-scale project highlighted the difficulty of collecting high-quality crowdsourced data, so I then studied how to refine crowdsourced data collection based on MTurk workers' own experiences. I found that MTurk workers often lie about their demographics due to privacy concerns and have strong opinions on issues such as fair quality filtering, often at odds with received wisdom in NLP; as a result, I proposed best practices for issues such as payment, quality checks, and considering worker incentives [Huang et al., 2023 - EMNLP Outstanding Paper].

LMs introduce a further challenge: language *itself* becomes an axis of discrimination. Disparate treatment based on features like dialect and accent allows harms to fly under the radar. I found that people ascribe consistent demographic personas and matching stereotypes to text-to-speech voices (despite no such marketing); e.g., a voice heard as older and feminine was rated as less competent, and those heard as older, masculine, and Black were rated as unfriendly [Fleisig et al., 2025]. Moreover, the voices' GitHub use cases entrench those stereotypes: e.g., the voice most often heard as feminine and Black was used for teaching tools and adult content generation. I also found that models like ChatGPT reinforce discrimination based on the user's dialect, often replying with either exaggerated dialect features or condescendingly formal responses [Fleisig et al., 2024]. This work has been picked up both within and outside computer science; reported on in *Nature*; and cited across over 10 disciplines, from education to immunology [11, 12, 13].

Now that LM use is more widespread, I plan to expand my work on capturing societal impacts of language models to evaluate the long-term effects of LMs on their users. By analyzing longitudinal

and larger-scale usage data, we can understand when and how long-term language model use affects users' worldviews, and whether content is served in discriminatory ways to populations of different backgrounds. Here, my long-term research agenda is to use longitudinal, large-scale harm evaluation to ground language model improvements in measurable real-world impact.

3 Deployment: Addressing model development issues to reduce downstream harms. When models are used by real-world communities, issues such as misuse and misinformation arise due to weaknesses in language model design. LMs are often confidently wrong; to help address this issue, I worked on creating GRACE, a framework for evaluating confidence calibration uniquely grounded in human performance on the same task [Sung*, Fleisig*, et al., 2025]. We leveraged a confidence calibration task that humans already do: a trivia format using long questions with increasingly easy clues that players compete to answer first, gauging when they are confident enough to answer. We trained models to compete with humans and collected live competition data, permitting granular model-human calibration comparison based on the speed, accuracy, and confidence of responses. GRACE reveals that language models are more *accurate* but worse *calibrated* than humans: unlike LMs, people tend to "know what they know." This framework also helps to improve adversarial datasets [Sung et al., 2024 - NAACL Outstanding Paper]. In addition, I worked on Ghostbuster, an AI-generated text detector that helps to catch issues such as students using AI to ghostwrite assignments or news stories generated with AI [Verma et al., 2024]. To do so, Ghostbuster introduces novel search techniques to detect AI-generated text. To mitigate ethical risks of misclassifying genuine work as AI-generated, Ghostbuster emphasizes robustness through evaluation on non-native English speakers' text and lightly edited AI-generated text. Ghostbuster was state-of-the-art when released, outperforming commercial closed-source models, and its demo website has over 500,000 hits.

Moving forward, I aim to examine the effects on overall human-AI performance when people collaborate with LMs in practice. For example, when models are used by human experts, such as doctors or lawyers making crucial decisions in their fields, how can we adjust language model confidence to account for their relative expertise? In ongoing work, I'm examining how people with different levels of expertise across domains collaborate with language models to answer questions, and whether people can be misled by confidently wrong explanations despite knowing the correct answer. My long-term agenda in this area is to design processes by which human-AI collaboration in high-stakes settings solves users' problems, rather than creating new ones.

Despite increasing concern over making LMs work for users with different needs, there are still further issues to ameliorate—issues that only become more pressing as LMs are used in an increasing range of complex settings. In future research, I aim to address these challenges: training models that meet different users' needs, evaluating complex harms to ground LM improvements in measurable real-world impact, and improving human-AI collaboration in high-stakes deployment settings. For each direction, I plan to collaborate broadly with researchers interested in NLP, AI ethics/fairness, and ML evaluation, to continue building language models that account for users' diverse needs.

References

Jessica Dai & Eve Fleisig. 2024. Mapping Social Choice Theory to RLHF. In *Workshop on Reliable and Responsible Foundation Models.*

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. 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 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Eve Fleisig, Julian Vargo, Nikolai Schwarz, Veronica Grajeda, Abigail Roberts, Rhosean Asmah, Nicole Holliday. 2025. Perceptions of OpenAI's Whisper Text-to-Speech Technology. *Linguistic Society of America Annual Meeting*.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating Worker Perspectives into MTurk Annotation Practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Harbani Jaggi, Kashyap Coimbatore Murali, Eve Fleisig, and Erdem Biyik. 2024. Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21910–21917, Miami, Florida, USA. Association for Computational Linguistics.

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, and Jordan Lee Boyd-Graber. 2025. GRACE: A Granular Benchmark for Evaluating Model Calibration against Human Calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19586–19587, Vienna, Austria. Association for Computational Linguistics.

Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Is your benchmark truly adversarial? AdvScore: Evaluating Human-Grounded Adversarialness. 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 623–642, Albuquerque, New Mexico. Association for Computational Linguistics.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. 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 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.

[1] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (IUI '23). Association for Computing Machinery, New York, NY, USA, 2–18.

[2] McKenzie Sadeghi, Dimitris Dimitriadis, Lorenzo Arvanitis, Virginia Padovese, Giulia Pozzi, Sara Badilini, Chiara Vercellone et al. 2025. Tracking AI-enabled Misinformation. *Newsguard.*

[3] Olivia Sidoti, Eugenie Park, and Jeffrey Gottfried. 2025. About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023. *Pew Research Center.*

[4] Hua, Yining, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou et al. Large language models in mental health care: a scoping review. 2025. Current Treatment Options in Psychiatry 12, no. 1, pages 1-18.

[5] Gabriel Reuben Smith, Carolina Bello, Lalasia Bialic-Murphy, Emily Clark, Camille S. Delavaux, Camille Fournier de Lauriere, Johan van den Hoogen et al. 2024. Ten simple rules for using large language models in science. *PLOS Computational Biology* 20, no. 1.

[6] Kashmir Hill. 2025. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. *The New York Times.*

[7] Yewon Kim, Thanh-Long V. Le, Donghwi Kim, Mina Lee, and Sung-Ju Lee. 2025. Design Opportunities for Explainable AI Paraphrasing Tools: A User Study with Non-native English Speakers. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 1061-1083.

[8] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, no. 8028, pages 147-154.

[9] Caroline Allen. 2022. Facebook's Content Moderation Failures in Ethiopia. *Council on Foreign Relations*.

[10] 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 P*roceedings 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.

[11] Laura Vargas-Parada. 2025. Large language models are biased — local initiatives are fighting for change. *Nature.*

[12] Phuong-Anh Nguyen. Evaluating AI-Generated Language as Models for Strategic Competence in English Language Teaching. *IAFOR Journal of Education* 12, no. 3 (2024): 325-349.

[13] Malik Sallam, Kholoud Al-Mahzoum, Omaima Alshuaib, Hawajer Alhajri, Fatmah Alotaibi, Dalal Alkhurainej, Mohammad Yahya Al-Balwah, Muna Barakat, and Jan Egger. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC Infectious Diseases* 24, no. 1 (2024): 799.

[14] Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. 2024. Preference learning algorithms do not learn preference rankings. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (NIPS '24), Vol. 37. Curran Associates Inc., Red Hook, NY, USA, Article 3234, 101928–101968.

[15] Prasann Singhal, Tanya Goyal, Jiacheng Xu, & Greg Durrett. 2024. A Long Way to Go: Investigating Length Correlations in RLHF.